

全国统计教材编审委员会推荐使用教材（2003 年第 2 版）

# SPSS 统计分析

（第 4 版）

卢纹岱 主编

吴喜之 审校

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

## 内 容 简 介

《SPSS 统计分析（第4版）》是在前三版的基础上，根据读者的反馈意见重新编写的。全书内容以统计分析应用为主，简要介绍各种统计分析方法的基本思想和基本概念；详细叙述操作方法，每种分析方法均给出对应的例题，涉及各个领域。每个例题均从方法选择、数据文件结构、操作步骤和结果分析方面给予说明。本书保留前三版的统计分析方法，压缩基本操作内容，增加结合分析、频谱分析和函数应用等内容。为方便读者和减少篇幅，书中所有例题数据均按章节编号，并保存在所附的光盘中。为便于教学，本书另配有电子教案，向采纳本书作为教材的教师免费提供。

本书可作为高等院校统计计算课程的本科生和研究生教材，也适合于从事分析和决策的社会各领域各相关专业读者学习参考。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容

版权所有·侵权必究

### 图书在版编目（CIP）数据

SPSS 统计分析 / 卢纹岱主编. —4版. —北京：电子工业出版社，2010.4

全国统计教材编审委员会推荐使用教材

ISBN 978-7-121-10580-7

I. ①S… II. ①卢… III. ①统计分析—软件包, SPSS—高等学校—教材 IV. ①C819

中国版本图书馆 CIP 数据核字（2010）第 049197 号

策划编辑：杨丽娟

责任编辑：杨丽娟

印 刷：北京市顺义兴华印刷厂

装 订：三河市双峰印刷装订有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本：787×980 1/16 印张：46 字数：979 千字

版 次：2010 年 4 月第 1 次印刷

印 数：5 000 册 定价：59.80 元（含光盘 1 张）

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 [zltts@phei.com.cn](mailto:zltts@phei.com.cn)，盗版侵权举报请发邮件至 [dbqq@phei.com.cn](mailto:dbqq@phei.com.cn)。

服务热线：（010）88258888。

## 《SPSS for Windows 统计分析（第4版）》编委会

主 编： 卢纹岱

审 校： 吴喜之

副主编： 朱红兵      朱一力      何丽娟      沙 捷

编 委： 殷小川      梁 蕾      张泰昌      宋楚强

卢纹凯      张晨      陈冬      苏林

费青松      石国书      王 湛      贺芬兰

宋 峥      朱启钊      盖文红      郭 娟

石国书      解利辉      朱江华      张 跃

# 前 言

SPSS软件原名为Statistical Package for the Social Science，社会科学用统计软件包。2000年SPSS公司将其英文全称改为“Statistical Product and Service Solutions”，意为“统计产品与服务解决方案”，是一个组合式软件包。它集数据整理、分析过程、结果输出等功能于一身，是世界著名的统计分析软件之一。

在我们推出下一版本时，SPSS可能将更改其名，用“PASW”取代“SPSS”。PASW英文全称为“Predictive Analytics Software”，即预测分析软件。

SPSS 使用Windows的窗口方式展示各种管理数据和分析方法的功能，使用对话框展示各种功能选择项，清晰、直观、易学易用，涵盖面广。读者只要掌握一定的Windows操作技能和统计分析原理，就可以使用该软件为特定的科研工作服务。即使统计学水平有限，也可以使用系统默认项得到初步的分析结果，从而免去了编写程序的复杂工作。由于它具有强大的图形功能，使用该软件不但可以得到分析后的数字结果，还可以得到直观、清晰、漂亮的统计图，形象地显示对原始数据和分析结果的各种描述。

SPSS已经在我国的社会科学和自然科学的各个领域得到广泛应用并发挥了巨大作用。我们所编写的《SPSS 统计分析》第1、2、3版得到广大读者的厚爱，成为受读者欢迎的畅销书就是一个很好的证明。

在前三版的基础上，我们编写了《SPSS 统计分析（第4版）》。

根据 PASW 软件的发展和广大读者的要求，我们对原作进行了仔细的检查、修正与改写，并按照增加内容但不增加篇幅的原则做了如下的改动。

1. 本书软件操作内容适用于 SPSS 16.0 以上版本，兼顾 15.0 以下版本。
2. 对基本操作的内容进行了进一步压缩。
3. 随着应用统计学知识的普及，并根据读者要求，相对于上一个版本，本书增加了：
  - 随机变量及随机变量函数应用；
  - 时间日期函数及其应用；
  - 在时间序列分析中增加了频谱分析、自相关、互相关分析方法；
  - 为适应市场研究的读者需要，增加了对调查数据的结合分析，以及与之有关的正交实验设计、调查表格和卡片的生成、打印等统计分析方法和应用实例。

本书共分三大部分。

- 第1章至第3章主要介绍 SPSS 的基本操作、基本概念和操作环境的设置，以及利用软件的各种帮助功能自学的方法。



- 第4章至第19章主要介绍随机变量和分布函数的应用、日期时间的运算；描述统计方法和分析表格的生成方法。还详细介绍了均值比较与检验、方差分析（参数检验）、非参数检验、相关分析、回归分析、聚类分析、判别分析、因子分析、对应分析、尺度分析、结合分析、时间序列分析、多响应变量分析、生存分析。

- 第20章和第21章详尽地介绍各种统计图形的生成、编辑、修饰的方法。

为便于初学者和非统计学专业的读者学习，本书章节的编排有利于读者由浅入深地系统学习统计学知识和正确选择分析方法。每章均对统计分析方法的基本思想或基本概念做了深入浅出的介绍；对软件的操作进行尽量详细的说明；并对每种分析方法配以相应的例题。本书各章节的例题从数据解释、数据文件结构、方法选择、软件操作、输出结果解释和结论等几方面加以详细的说明。本书大部分例题均为作者科研或教学中的实例，读者容易接受。

为适应SPSS软件老用户的编程的需要，本版仍保留了语句部分，作为包括语句部分的最后一个版本。

本书所有例题数据按章节编号保存在本书所附的光盘中，数据文件名均以“data”开头，接着是2位数字的章号，横线后是2位数字，表明数据文件在本章中出现的序号。文件类型主要是SPSS数据文件（dataxx-xx.sav），也有少量Excel文件（dataxx-xx.xls）和文本文件（dataxx-xx.txt）。读者可以按照书中的数据图例查找并参照。为读者学习方便，每个分析方法的介绍除有些基本操作被简化外，基本彼此独立，读者可根据自己的需要自行安排阅读。

本书由卢纹岱主编，并特别邀请中国人民大学统计学院吴喜之教授审校，在此深表谢意！

本书各章编写情况如下：卢纹岱、张泰昌、宋峥完成了第1～5章；宋楚强、郭娟、张跃完成了第8～10章；卢纹岱、陈冬共同完成第13章；第16章由卢纹岱、朱江华完成；第18章由卢纹岱、张晨、梁蕾完成；第6、12、17章由朱红兵、苏林、朱启钊共同完成；第19～21章由朱一力、王湛、贺芬兰编写；何丽娟、崔健完成了第7章；沙捷、盖文红共同完成了第11章，卢纹岱、朱红兵合写了第14章，何丽娟与殷小川合写了第15章。全书的统稿及排版工作由卢纹岱负责。在编写过程中，金水高、卢纹凯、张泰昌教授、梁蕾副教授提供了部分例题数据。石国书、解利辉、费青松、王雁等老师在资料收集、数据录入、核对、利用SPSS软件绘图等方面做了大量工作，在此一并表示诚挚的感谢。

本书适用于从事数据分析或统计应用的各领域、各专业的研究人员、中高层管理人员和决策者，也可以作为要求掌握统计分析方法和SPSS软件操作的高等院校的本科生、研究生的教材和自学参考书。

为方便教学，本书另配有电子教案，向采纳本书作为教材的教师免费提供，可登录电子工业出版社华信教育资源网（[www.hxedu.com.cn](http://www.hxedu.com.cn)）或电话联系（010-88254537）获取。

由于水平有限，加之时间仓促，有待改进的地方仍然很多，不妥之处在所难免，恳请广大读者对本书继续提出批评指正，我们愿与各位同行和爱好者进行交流学习。反馈意见请发电子邮件至：

[luwendai@tsinghua.org.cn](mailto:luwendai@tsinghua.org.cn)

[zhuhongbing@cipe.net.cn](mailto:zhuhongbing@cipe.net.cn)

[zhuyili2008@sina.com](mailto:zhuyili2008@sina.com)

[helijuan@cipe.net.cn](mailto:helijuan@cipe.net.cn)

[shajie@cipe.net.cn](mailto:shajie@cipe.net.cn)

编 者

# 目 录

第 1 章 SPSS 概述	1
1.1 软件安装与运行	1
1.1.1 SPSS 软件安装方法	1
1.1.2 SPSS 启动与退出	1
1.1.3 SPSS 运行管理方式	3
1.2 窗口及其功能概述	4
1.2.1 数据编辑窗口	4
1.2.2 输出窗口	5
1.2.3 语句窗口	6
1.2.4 窗口菜单	8
1.2.5 对话框及其使用方法	8
1.2.6 设置工具栏中的工具图标按钮	10
1.3 系统参数设置	12
1.3.1 参数设置基本操作	12
1.3.2 通用参数设置	13
1.3.3 输出观察窗口参数设置	14
1.3.4 数据属性参数设置	15
1.3.5 货币变量自定义格式设置	17
1.3.6 标签输出设置	17
1.3.7 统计图形参数设置	18
1.3.8 输出表格参数设置	22
1.3.9 文件默认存取位置设置	23
1.4 统计分析功能概述	24
1.5 数据与变量	25
1.5.1 常量与变量	25
1.5.2 操作符与表达式	28
1.5.3 概率事件	30
1.5.4 SPSS 函数	30
1.6 获得帮助	37
1.6.1 SPSS 帮助系统	37
1.6.2 右键帮助	40

习题 1.....	41
<b>第 2 章 数据与数据文件.....</b>	<b>42</b>
2.1 变量定义与数据编辑.....	42
2.1.1 数据编辑器.....	42
2.1.2 定义变量.....	43
2.1.3 定义日期变量.....	46
2.1.4 数据录入与编辑.....	48
2.1.5 根据已有的变量建立新变量.....	52
2.1.6 建立值标签的工具与程序.....	54
2.1.7 打开、保存与查看数据文件.....	57
2.2 数据文件的转换.....	60
2.2.1 ASCII 码数据文件的转换.....	60
2.2.2 数据库文件的转换.....	67
2.2.3 观测量的查重.....	72
2.3 数据文件操作.....	75
2.3.1 数据文件的拆分与合并.....	75
2.3.2 观测量的排序与排秩.....	81
2.3.3 对变量值重新编码.....	83
2.3.4 数据文件的转置与重新构建.....	87
2.4 观测量的加权与选择.....	100
2.4.1 定义加权变量.....	100
2.4.2 选择参与分析的观测量.....	101
习题 2.....	102
<b>第 3 章 输出信息的编辑.....</b>	<b>104</b>
3.1 输出窗口中的文本浏览与编辑.....	104
3.1.1 利用导航器浏览输出信息.....	104
3.1.2 编辑导航器中的输出项.....	106
3.2 输出表格中信息的编辑.....	107
3.2.1 表格编辑工具与常用编辑方法.....	107
3.2.2 表格的转置与行、列、层的处理.....	109
3.2.3 表格外观的设置与编辑.....	112
3.2.4 输出信息的复制与打印.....	116
习题 3.....	116

第 4 章 随机变量与分布函数的应用	117
4.1 随机变量与分布函数	117
4.1.1 随机变量及其概率分布	117
4.1.2 随机变量函数	121
4.2 随机变量与分布函数应用	128
4.2.1 符合分布要求的随机数的生成	128
4.2.2 概率密度函数与累积概率密度函数的应用	130
习题 4	133
第 5 章 日期和时间函数及其运算	134
5.1 日期时间函数	134
5.1.1 SPSS 日期时间概述	134
5.1.2 日期时间常量与变量	134
5.1.3 日期时间函数	137
5.2 日期时间函数的应用	140
5.2.1 日期时间型变量的格式转换	140
5.2.2 日期时间型变量的算术运算	143
习题 5	145
第 6 章 构建表格	146
6.1 自定义表格	146
6.1.1 自定义表格的概念	146
6.1.2 自定义表格的操作	147
6.2 汇总、统计指标与统计检验	149
6.2.1 统计指标与汇总项	149
6.2.2 表格中的统计检验	154
6.3 标题与其他选项	154
6.3.1 定义表格标题	154
6.3.2 定义表格选项	155
6.4 自定义表格实例	156
6.5 自定义表格的过程语句	157
习题 6	161
第 7 章 基本统计分析	162
7.1 频数分布分析	162

7.1.1 频数分布分析过程 .....	162
7.1.2 频数分布分析实例 .....	165
7.2 描述统计 .....	166
7.2.1 描述统计中的基本概念 .....	167
7.2.2 描述统计分析过程 .....	167
7.2.3 描述统计分析实例 .....	168
7.3 探索分析 .....	169
7.3.1 探索分析的意义和数据要求 .....	169
7.3.2 探索分析过程 .....	171
7.3.3 探索分析实例 .....	173
7.4 列联表分析 .....	176
7.4.1 列联表及其独立性卡方检验的思路 .....	176
7.4.2 列联表分析过程 .....	177
7.4.3 列联表分析实例 .....	181
7.5 比率分析 .....	183
7.5.1 比率分析过程 .....	184
7.5.2 比率分析实例 .....	185
7.6 P-P 图和 Q-Q 图 .....	186
7.6.1 P-P 图和 Q-Q 图分析过程 .....	186
7.6.2 P-P 图和 Q-Q 图分析实例 .....	187
习题 7 .....	190

第 8 章 均值比较与检验 .....	191
8.1 均值比较与均值比较的检验 .....	191
8.1.1 均值比较的概念 .....	191
8.1.2 均值比较与检验的过程 .....	191
8.2 MEANS 过程 .....	193
8.2.1 MEANS 过程中的统计量 .....	193
8.2.2 MEANS 过程操作 .....	195
8.2.3 分析实例 .....	196
8.2.4 MEANS 过程语句 .....	200
8.3 单一样本 T 检验 .....	202
8.3.1 单一样本 T 检验的概念 .....	202
8.3.2 单一样本 T 检验的实例 .....	202
8.4 独立样本 T 检验 .....	204

8.4.1	独立样本 T 检验的概念 .....	204
8.4.2	独立样本 T 检验的过程 .....	205
8.4.3	独立样本 T 检验的实例 .....	206
8.5	配对样本 T 检验 .....	209
8.5.1	配对样本 T 检验的概念 .....	209
8.5.2	配对样本 T 检验的过程 .....	210
8.5.3	配对样本 T 检验的实例 .....	210
8.6	T 检验过程语句 .....	212
习题 8	.....	214
<b>第 9 章</b>	<b>方差分析</b> .....	<b>215</b>
9.1	方差分析的概念与方差分析过程 .....	215
9.1.1	方差分析的概念 .....	215
9.1.2	方差分析中的术语 .....	217
9.1.3	方差分析过程 .....	219
9.2	单因素方差分析 .....	220
9.2.1	简单的一维方差分析 .....	221
9.2.2	单因素方差分析过程 .....	223
9.2.3	单因素方差分析实例 .....	226
9.2.4	单因素方差分析过程语句 .....	231
9.3	单因变量多因素方差分析 .....	233
9.3.1	单因变量多因素方差分析概述 .....	233
9.3.2	单因变量多因素方差分析过程 .....	234
9.3.3	随机区组设计的方差分析实例 .....	240
9.3.4	2×2 析因实验方差分析实例 .....	242
9.3.5	拉丁方区组设计的方差分析实例 .....	246
9.3.6	协方差分析实例 .....	249
9.3.7	多维交互效应方差分析实例 .....	251
9.4	多因变量线性模型的方差分析 .....	254
9.4.1	多因变量方差分析概述 .....	254
9.4.2	多因变量方差分析过程和数据要求 .....	255
9.4.3	多因变量线性模型方差分析实例 .....	257
9.5	重复测量设计的方差分析 .....	269
9.5.1	重复测量方差分析概述 .....	269
9.5.2	重复测量方差分析的数据文件结构 .....	272

9.5.3	组内因素的设置与重复测量方差分析过程 .....	273
9.5.4	重复测量方差分析实例 .....	276
9.5.5	关于趋势分析 .....	279
9.6	方差成分分析 .....	283
9.6.1	方差成分分析过程 .....	284
9.6.2	方差成分分析实例 .....	287
习题 9	.....	290
<b>第 10 章</b>	<b>相关分析</b> .....	<b>292</b>
10.1	相关分析的概念与相关分析过程 .....	292
10.1.1	简单相关分析的概念 .....	292
10.1.2	相关分析过程 .....	293
10.2	两个变量间的相关分析 .....	294
10.2.1	两变量间相关分析过程 .....	294
10.2.2	两个变量间相关分析实例 .....	295
10.2.3	两个变量相关分析的过程语句 .....	299
10.2.4	关于相关矩阵 .....	301
10.2.5	建立相关矩阵数据文件 .....	302
10.3	偏相关分析 .....	305
10.3.1	偏相关分析的概念 .....	305
10.3.2	偏相关分析过程 .....	306
10.3.3	偏相关分析实例 .....	307
10.3.4	偏相关分析的过程语句 .....	309
10.4	距离分析 .....	311
10.4.1	距离分析的概念 .....	311
10.4.2	距离分析过程 .....	312
10.4.3	距离分析实例 .....	314
习题 10	.....	317
<b>第 11 章</b>	<b>回归分析</b> .....	<b>318</b>
11.1	线性回归 .....	318
11.1.1	一元线性回归 .....	318
11.1.2	多元线性回归 .....	320
11.1.3	异常值、影响点、共线性诊断 .....	322
11.1.4	变非线性关系为线性关系 .....	324



11.1.5 线性回归过程	325
11.1.6 线性回归分析实例	329
11.2 曲线估计	333
11.2.1 曲线回归概述	333
11.2.2 曲线回归过程	334
11.2.3 曲线回归分析实例	335
11.3 二项逻辑斯谛回归	337
11.3.1 Logistic 回归模型	337
11.3.2 二项逻辑斯谛回归过程	341
11.3.3 二项逻辑斯谛回归分析实例	343
11.4 多分变量的逻辑斯谛回归	347
11.4.1 多分变量逻辑斯谛回归的概念	347
11.4.2 多分变量的逻辑斯谛回归过程	349
11.4.3 多分变量逻辑斯谛回归分析实例	352
11.5 概率单位回归	356
11.5.1 概率单位回归的概念	356
11.5.2 概率单位回归过程	357
11.5.3 概率单位回归分析实例	359
11.6 非线性回归	361
11.6.1 非线性模型	361
11.6.2 非线性回归过程	364
11.6.3 非线性回归分析实例	366
11.7 加权回归	369
11.7.1 加权回归的概念	369
11.7.2 加权回归过程	370
11.7.3 加权回归分析实例	371
习题 11	373

第 12 章 非参数检验	374
12.1 卡方检验	374
12.1.1 卡方检验的基本概念	374
12.1.2 卡方检验过程	375
12.1.3 卡方检验分析实例	376
12.2 二项分布检验	378
12.2.1 二项分布检验的概念与操作	378

12.2.2 二项分布检验分析实例 .....	379
12.3 游程检验 .....	380
12.3.1 游程检验的基本概念 .....	380
12.3.2 游程检验过程 .....	381
12.3.3 游程检验分析实例 .....	381
12.4 一个样本的柯尔莫哥洛夫-斯米诺夫检验 .....	382
12.4.1 一个样本的柯尔莫哥洛夫-斯米诺夫检验的基本概念 .....	382
12.4.2 柯尔莫哥洛夫-斯米诺夫检验过程 .....	383
12.4.3 柯尔莫哥洛夫-斯米诺夫检验分析实例 .....	383
12.5 两个独立样本检验 .....	384
12.5.1 两个独立样本检验的用途与基本操作 .....	384
12.5.2 两个独立样本检验分析实例 .....	385
12.6 多个独立样本检验 .....	386
12.6.1 多个独立样本检验的用途与操作 .....	386
12.6.2 多个独立样本检验分析实例 .....	387
12.7 两个相关样本检验 .....	388
12.7.1 两个相关样本检验的用途与操作 .....	388
12.7.2 两个相关样本检验分析实例 .....	389
12.8 多个相关样本检验 .....	389
12.8.1 多个相关样本检验的用途与操作 .....	389
12.8.2 多个相关样本检验分析实例 .....	390
12.9 非参数假设检验过程的命令语句 .....	391
习题 12 .....	396
 第 13 章 聚类分析与判别分析 .....	 397
13.1 聚类、判别分析及其分析过程 .....	397
13.1.1 聚类分析 .....	397
13.1.2 判别分析 .....	398
13.1.3 聚类与判别分析过程 .....	398
13.2 两步聚类 .....	398
13.2.1 两步聚类概述 .....	398
13.2.2 两步聚类过程 .....	400
13.2.3 两步聚类分析实例 .....	404
13.2.4 两步聚类过程的命令语句 .....	409
13.3 快速样本聚类 .....	412

13.3.1	快速样本聚类概述 .....	412
13.3.2	快速样本聚类过程 .....	413
13.3.3	快速样本聚类分析实例 .....	416
13.3.4	快速样本聚类过程的命令语句 .....	420
13.4	分层聚类 .....	422
13.4.1	分层聚类概述 .....	422
13.4.2	分层聚类过程 .....	423
13.4.3	样本分层聚类分析实例 .....	428
13.4.4	变量聚类概述 .....	436
13.4.5	变量聚类分析实例 .....	437
13.4.6	分层聚类过程的命令语句 .....	441
13.5	判别分析 .....	445
13.5.1	判别分析概述 .....	445
13.5.2	判别分析过程 .....	447
13.5.3	判别分析实例 .....	452
13.5.4	逐步判别分析与实例 .....	461
13.5.5	判别分析过程的命令语句 .....	467
习题 13	.....	471
第 14 章	因子分析与对应分析 .....	472
14.1	主成分分析与因子分析 .....	472
14.1.1	主成分分析与因子分析概述 .....	472
14.1.2	因子分析过程 .....	478
14.1.3	因子分析实例 .....	484
14.1.4	利用因子得分进行聚类 .....	488
14.1.5	市场研究中的顾客偏好分析 .....	492
14.1.6	因子分析过程的命令语句 .....	496
14.2	对应分析 .....	500
14.2.1	对应分析概述 .....	500
14.2.2	对应分析过程 .....	501
14.2.3	对应分析实例 .....	504
14.2.4	对应分析过程的命令语句 .....	506
习题 14	.....	509

第 15 章 尺度分析	510
15.1 信度分析	510
15.1.1 信度分析的概念	510
15.1.2 信度分析过程	513
15.1.3 信度分析实例	515
15.2 多维尺度分析 (ALSCAL)	517
15.2.1 多维尺度分析的功能与数据要求	517
15.2.2 多维尺度分析过程	517
15.2.3 多维尺度分析实例	520
习题 15	522
第 16 章 结合分析	523
16.1 结合分析概述	523
16.2 正交实验设计	524
16.2.1 实验设计中的问题	524
16.2.2 正交实验设计的思路	525
16.2.3 正交实验设计过程	526
16.2.4 正交实验设计实例	529
16.2.5 正交设计过程语句	530
16.3 实验设计结果打印	535
16.3.1 设计结果打印过程	535
16.3.2 打印调查用卡片实例	536
16.3.3 正交实验设计打印过程语句	537
16.4 结合分析的语句与编程	539
16.4.1 结合分析过程语句	539
16.4.2 结合分析语句实例	544
16.5 结合分析实例	548
16.5.1 课题分析与正交设计	548
16.5.2 调查准备与调查	550
16.5.3 结合分析编程与结果分析	552
习题 16	556
第 17 章 时间序列分析	557
17.1 时间序列的建立和平稳化	558
17.1.1 缺失值数据的修补	558

17.1.2 建立时间序列新变量 .....	559
17.2 序列图 .....	561
17.2.1 序列图过程 .....	561
17.2.2 序列图应用实例 .....	563
17.3 建立时间序列模型 .....	564
17.3.1 指数平滑与 ARIMA 模型概述 .....	564
17.3.2 选择分析变量 .....	565
17.3.3 选择统计量 .....	571
17.3.4 Plots 图形 .....	573
17.3.5 输出项目的过滤 .....	574
17.3.6 保存新变量 .....	574
17.3.7 建模的其他选择项 .....	575
17.3.8 时间序列分析实例 .....	576
17.4 应用时间序列模型 .....	580
17.4.1 应用时间序列模型过程 .....	580
17.4.2 应用时间序列模型分析实例 .....	581
17.5 自相关 .....	582
17.5.1 自相关图 .....	582
17.5.2 自相关分析过程 .....	583
17.5.3 自相关分析实例 .....	584
17.6 季节分解法 .....	585
17.6.1 季节分解法分析过程 .....	586
17.6.2 季节分解法分析实例 .....	586
17.7 频谱分析 .....	587
17.7.1 频谱分析概述 .....	587
17.7.2 频谱分析过程 .....	588
17.7.3 频谱分析实例 .....	589
17.8 互相关 .....	590
17.8.1 互相关概述 .....	590
17.8.2 互相关过程 .....	591
17.8.3 互相关实例 .....	591
习题 17 .....	592

第 18 章 多响应变量的分析 .....	593
-----------------------	-----

18.1 多响应变量的概念与分类 .....	593
------------------------	-----

18.2	定义与建立多响应变量集	595
18.3	多响应变量的频数分布分析	596
18.3.1	多响应二分变量集的频数分布分析	596
18.3.2	多响应分类变量集的频数分布分析	597
18.4	多响应变量的交叉表分析	601
18.4.1	多响应变量集交叉表分析过程	601
18.4.2	多响应二分变量集的交叉表分析实例	602
18.5	多响应变量集分析的过程语句	605
18.6	使用 Table 功能分析多响应变量集	608
18.6.1	简单频数分布分析	609
18.6.2	交叉表分析	610
习题 18		613
<b>第 19 章</b>	<b>生存分析</b>	<b>614</b>
19.1	生存分析概述	614
19.1.1	生存分析与生存数据	614
19.1.2	生存时间函数	614
19.1.3	Cox 回归模型	615
19.2	生命表分析	615
19.2.1	生命表分析概述	615
19.2.2	生命表分析过程	616
19.2.3	生命表分析实例	617
19.3	Kaplan-Meier 分析	621
19.3.1	Kaplan-Meier 分析概述	621
19.3.2	Kaplan-Meier 分析过程	621
19.3.3	Kaplan-Meier 分析实例	623
19.4	Cox Regression 风险比例模型分析	625
19.4.1	Cox Regression 分析概述	625
19.4.2	Cox Regression 分析过程	625
19.4.3	Cox Regression 分析实例	629
习题 19		631
<b>第 20 章</b>	<b>生成统计图形</b>	<b>632</b>
20.1	概述	632
20.2	条形图和 3-D 条形图	633

20.2.1	选择图形类型 .....	633
20.2.2	观测量分组描述简单条形图 .....	634
20.2.3	变量模式简单条形图 .....	637
20.2.4	观测量分组模式分段条形图 .....	637
20.2.5	3-D 条形图 .....	638
20.3	线图、面积图和高低图 .....	639
20.3.1	选择图形类型 .....	639
20.3.2	观测值模式堆栈面积图 .....	640
20.3.3	观测量分类模式多线图 .....	641
20.3.4	变量模式垂线图 .....	642
20.3.5	观测量分类模式简单高低收盘图 .....	642
20.3.6	变量模式分组高低收盘图 .....	643
20.3.7	观测量分类模式简单极差图 .....	644
20.3.8	变量模式简单极差图 .....	644
20.3.9	观测值分类分组极差图 .....	646
20.3.10	变量模式差分面积图 .....	646
20.4	圆图 .....	647
20.4.1	观测量分类模式圆图 .....	647
20.4.2	变量模式圆图 .....	648
20.4.3	观测值模式圆图 .....	648
20.5	箱图和误差条图 .....	649
20.5.1	选择箱图和误差条图类型 .....	649
20.5.2	观测量分类模式简单箱图 .....	650
20.5.3	观测量分类模式简单误差条图 .....	650
20.5.4	变量模式简单箱图 .....	651
20.5.5	观测量分类模式分组误差条图 .....	651
20.5.6	变量模式分组箱图 .....	652
20.6	散点图 .....	653
20.6.1	选择散点图图式 .....	653
20.6.2	简单散点图 .....	654
20.6.3	重叠散点图 .....	654
20.6.4	矩阵散点图 .....	655
20.6.5	三维散点图 .....	656
20.6.6	简单点图 .....	656
20.7	直方图 .....	657

20.8 交互图 .....	658
20.8.1 交互式条形图、点图、线图和面积图 .....	658
20.8.2 交互式圆图 .....	661
20.8.3 交互式箱图和误差条图 .....	663
20.8.4 交互式直方图 .....	664
20.8.5 交互式散点图 .....	665
20.9 帕累托图 .....	667
20.9.1 选择帕累托图类型 .....	667
20.9.2 观测量分类数目或数值累加模式简单帕累托图 .....	667
20.9.3 变量累加模式简单帕累托图 .....	668
20.9.4 观测值模式简单帕累托图 .....	669
20.9.5 观测量数目或数值累加模式堆栈帕累托图 .....	669
20.9.6 变量累加模式堆栈帕累托图 .....	670
20.9.7 观测值模式堆栈帕累托图 .....	671
20.10 控制图 .....	671
20.10.1 选择控制图类型 .....	671
20.10.2 观测量组结构的平均值、极差、标准差控制图 .....	672
20.10.3 观测量组结构的单值-移动极差控制图 .....	673
20.10.4 观测量组结构数据的不合格品率、不合格品数控制图 .....	673
20.10.5 观测量组结构的缺陷数、单位缺陷数控制图 .....	674
20.10.6 变量组结构数据的平均值、极差、标准差控制图 .....	675
20.10.7 变量组结构数据的不合格品率、不合格品数控制图 .....	675
习题 20 .....	677

第 21 章 编辑统计图形 .....	678
21.1 认识图形组成 .....	678
21.2 编辑平面统计图 .....	679
21.2.1 图形编辑途径 .....	679
21.2.2 改变图形构成 .....	680
21.2.3 图形与文字修饰 .....	684
21.2.4 坐标轴的编辑 .....	686
21.2.5 图条的修饰 .....	689
21.2.6 图线的编辑 .....	690
21.2.7 圆图编辑 .....	692
21.2.8 散点图的编辑 .....	693



21.2.9 文件管理 .....	696
习题 21 .....	697
附录 A 标准化、距离和相似性的计算 .....	698
附录 B 数据清单 .....	705
参考文献 .....	711

# 第 1 章 SPSS 概 述

## 1.1 软件安装与运行

### 1.1.1 SPSS软件安装方法

(1) 开机，启动 Windows，将 SPSS 系统安装光盘放入光盘驱动器。

(2) 启动 Windows 资源管理器，双击光盘驱动器图标，在目录窗口中找到安装程序图标，如图 1-1(a)所示，两个安装图标均可。双击其中任意一个图标启动 SPSS 16.0，见图 1-1(b)，自动转入安装程序的屏幕显示如图 1-1(c)所示。

(3) 单击 Next 按钮，系统自动进行软件包的解压缩工作，安装开始以后就可以按照屏幕提示一步步地进行操作，每步操作均要认真阅读屏幕显示的信息和提示。

当再次出现如图 1-1(c)所示画面，且按钮区出现 Finish 按钮时，则单击该按钮，结束安装。

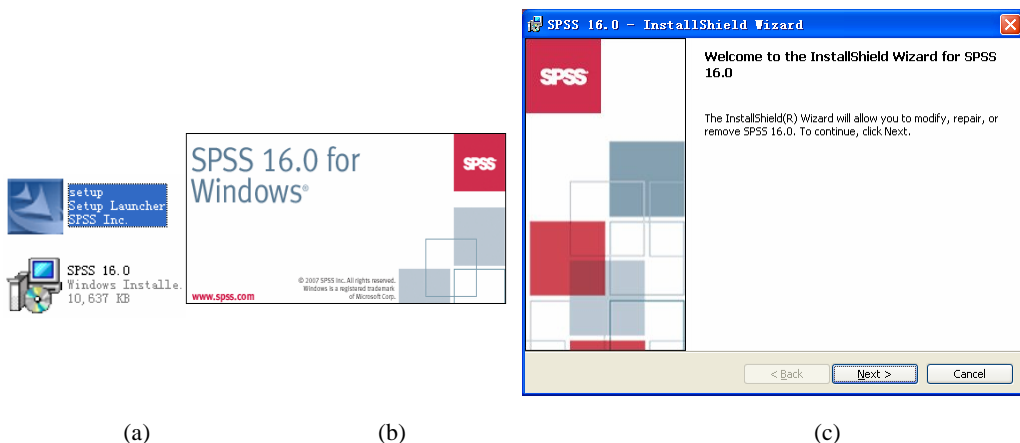


图 1-1 SPSS 的安装画面

### 1.1.2 SPSS启动与退出

#### 1. SPSS 的启动

(1) 开机后，启动了 Windows，双击 SPSS 的图标，将出现版本提示画面，如图 1-2(a)

所示。

(2) 在提示画面后出现 SPSS 文件选择对话框，共有 6 个功能选项和一个复选项，如图 1-2(b)所示。功能选项如下：

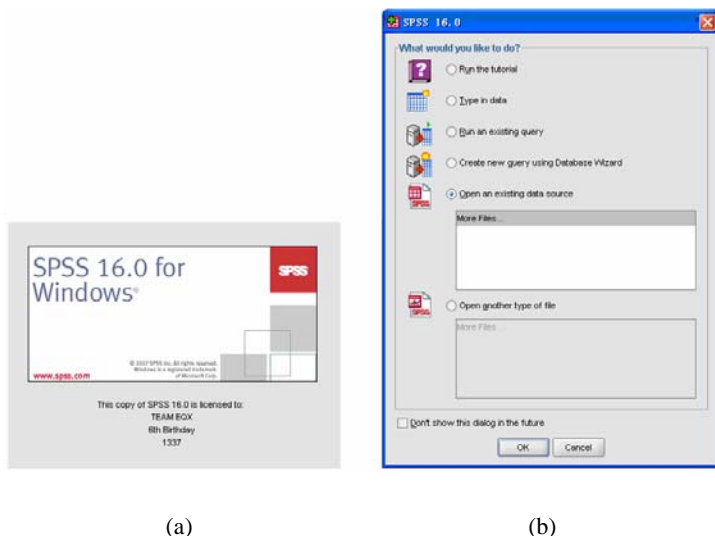


图 1-2 版本提示画面和文件选择对话框

• **Run the tutorial** 运行操作指南。选择此项打开如图 1-3(a)所示的操作指南。可以根据主题单击书形图标，查看基本操作指导信息。

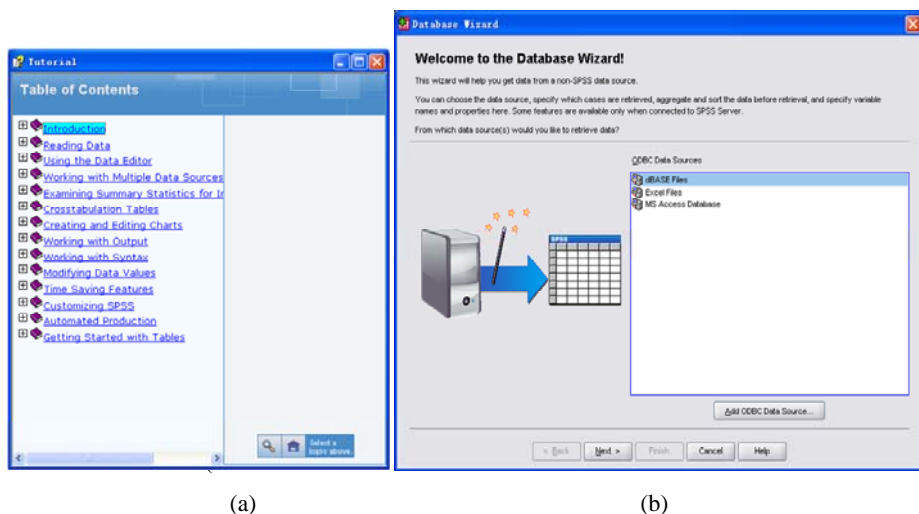


图 1-3 打开 SPSS 操作指南

• **Type in data** 在数据窗口中输入数据。选择此项则显示数据编辑窗口，等待输入数据建立新数据集。

• **Run an existing query** 运行现有的数据文件。选择此项将显示打开文件窗口，让读者选择一个\*.spq 文件。

• **Create new query using Database Wizard** 使用数据库创建新的数据文件。选择此项打开如图 1-3(b)所示的数据库处理工具，将诸如 DBF 格式文件、XLS 格式的 Excel 文件，以及 SQL 等数据库文件转换成 SPSS 数据文件。数据库处理工具使用方法参见第 2 章。

• **Open an existing data source** 打开已经存在的数据文件。选择此项读者可在第一个文件栏中选择一个.sav 格式的 SPSS 数据文件。

• **Open another type of file** 打开另一类型的文件。此项将让读者在第二个文件栏中选择一个其他格式的文件。例如选择一个常用的\*.spv，即 SPSS 的输出文件等。

如果在提示画面上勾选 **Don't show this dialog in the future** 项，下次启动 SPSS 时将不显示该对话框，直接显示空数据编辑窗口。

(3) 在对话框中单击 **Cancel** 按钮，跳过上述各项的选择，显示空数据编辑窗口 **SPSS Data Editor**，直接进入数据编辑状态，可以直接输入数据或操作菜单打开已经存在的数据文件。

## 2. SPSS 的退出

以下方法均可以达到退出 SPSS 系统的目的：

(1) 双击主画面左上角的窗口控制菜单图标，或单击该图标，在展开的小菜单中，单击“关闭”菜单项。

(2) 单击主菜单的 **File** 菜单项，在展开的文件菜单中，单击 **Exit** 命令。

(3) 单击数据编辑窗口右上角的  图标。

## 1.1.3 SPSS运行管理方式

### 1. 完全窗口菜单运行管理方式

SPSS 启动后即在屏幕上显示主画面，即数据编辑窗口，如图 1-4 所示。完全窗口菜单管理方式是从数据输入、编辑、分析一直到分析结果的打印输出都在窗口中显示，通过菜单、对话框操作进行。

完全窗口运行管理方式主要在数据编辑窗口和输出窗口中进行操作。这种运行方式操作简便、直观，特别适用于初学者。由于窗口中包括的是基本参数和基本统计量的选项，因此完全窗口运行管理方式对某些专业人员来说，可能不能充分满足

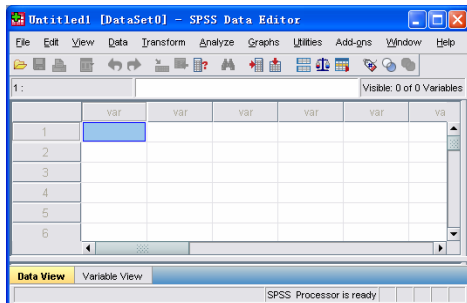


图 1-4 SPSS 数据编辑窗口

需要。

## 2. 程序运行管理方式

程序运行管理方式是在 **Syntax**（语句）窗口中直接运行编写好的程序的一种方式。在该窗口中输入由 **SPSS** 命令组成的程序，利用键盘或主菜单中 **Edit** 菜单项对窗口中的程序进行修改、编辑。在 **Syntax** 窗口中的程序可以分析数据窗口中的数据，也可以用有关的语句指定外部数据文件，对其进行分析，分析结果仍然显示在输出窗口中。对于习惯使用 **SPSS** 语言编写程序的读者仍然有用武之地。

## 3. 混合运行管理方式

混合运行管理方式是以上两种方法的结合方式。首先在数据窗口中输入数据或利用 **File** 菜单项打开已经存在的数据文件，然后利用对话框选择分析过程和分析参数。选择完成后不立即执行，而是用 **Paste** 按钮将选择的过程及参数转换成相应的命令语句，置于 **Syntax** 窗口中。在该语句窗口中增加对话框中没有包括的语句和参数，或修改子命令中的参数，然后单击窗口中的“运行”功能按钮，将程序提交系统执行，结果显示在输出窗口中。混合运行管理方式，既能简化操作，又可以弥补单纯窗口运行管理方式的不足。对于要求较高的统计分析功能，通常可使用这种方式。

# 1.2 窗口及其功能概述

**SPSS** 的文件系统包括 4 种基本类型的文件：**Data**（数据文件）、**Syntax**（语句文件）、**Output**（输出文件）和 **Script**（程序编辑文件）。每种类型的文件在各自的窗口中通过各自的菜单、功能按钮实现自己的各项功能。系统菜单的 **File** 下拉菜单中的 **New** 命令主要针对 4 个窗口中文件的操作，即当鼠标单击 **File** 菜单中 **New** 命令打开小菜单，显示可以新建各种类型的文件：**Data**、**Syntax**、**Output**、**Script**。对于使用 **SPSS** 的统计分析功能的读者来说，主要使用 3 种窗口，即数据窗口、输出窗口和语句窗口。

## 1.2.1 数据编辑窗口

**SPSS** 系统启动后激活该数据编辑窗口，如图 1-4 所示。未命名的数据编辑窗口最上方标有“**Untitled  $n$  [Data set  $m$ ] - SPSS Data Editor**”， $n$ 、 $m$  是打开窗口或数据文件的顺序号。窗口中有一个可扩展的平面二维表格，可以在此窗口中编辑数据文件。一旦保存了数据窗口中的数据，标题栏则显示该数据文件名。

对于数据窗口来说，无论“新建”还是“打开”命令，都会建立一个新的数据窗口。一次启动 **SPSS** 可以同时打开两个或两个以上的数据窗口。便于同时查看、操作两个以上的数据文件。单击标题栏激活数据编辑窗。被激活的数据编辑窗口标题栏为蓝色（默认），是当前工作窗；未被激活的数据编辑窗标题栏是灰色的。

## 1.2.2 输出窗口

SPSS 输出窗口标题栏中标有“Output1 [Document1] - SPSS Viewer”，按照 SPSS 默认设置，输出窗口在启动后不显示在屏幕的主画面上。

1. 使用以下方法可以使输出窗口激活并显示在屏幕画面上

(1) 当使用了 **Analyze** 菜单中的统计分析功能处理数据窗口中的数据产生输出信息时，输出窗口自动激活，显示在屏幕画面上。如果处理成功，则显示分析结果；如果处理过程中无法运行或发生错误，则在该窗口中显示系统给出的错误信息。

(2) 在 **File** 菜单中选择 **New** 项，在二级菜单中选择 **Output** 项，屏幕画面上显示一个输出窗口，如图 1-5 所示。可以同时打开几个输出窗口，在窗口最上方的标题栏中按打开顺序显示窗口名：**Output1**、**Output2**、**Output3**……，在保存输出内容时由读者给出具体名称。

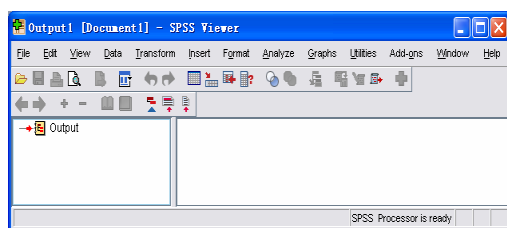


图 1-5 手动激活输出窗口

### 2. 输出窗口组成

输出窗口除标题栏外，还包括以下几部分。

- 主菜单：由 **File**~**Help**13 个菜单项组成。
- 工具栏：由各种功能的图标组成，是各种常用功能命令的快捷操作方式。
- 输出文本窗口：图标行下面右半边是一个文本窗口，在执行指定的操作或分析程序后，该窗口被激活，窗口显示输出信息，包括输出标题、文本、表格和统计图。该窗口中的内容可以利用鼠标、键盘和 **Edit** 菜单项的各种功能进行编辑。
- 输出导航窗口：导航窗口是浏览输出信息的导航器，位于图标行下面的左半边。以树形结构给出输出信息的提纲。
- 状态行：输出窗口的最下面一行是状态行，分为 5 个区，用鼠标指向任意一个区，就会在最左面区域显示每个区的功能解释。

### 3. 多个输出窗口的建立与主窗口的概念


利用 **File** 菜单中的 **New** 命令可以再打开一个输出窗口，打开的输出窗口按先后顺序标有 **Output2**、**Output3**……过程执行结果只会输出到当前窗口即工作输出窗口。标记为工具栏中的灰色十字。其他输出窗口，非当前输出窗口标记为工具栏中的绿色十字。

工作输出窗口（或称当前输出窗口）只能有一个。鼠标光标移到一个输出窗口中，单击该输出窗口中的十字图标按钮，就把该输出窗口激活为工作窗口。所有操作只对被激活的窗口有效。非激活窗口的窗口工具栏中的十字图标是绿色的。

**File** 菜单中的 **New** 命令可以打开一个新的空输出窗口，**Open** 命令把已经存在的输出文件显示到激活的输出窗口中。该输出窗口自动成为当前工作窗口。窗口标题栏显示

文件名。

#### 4. 关闭输出窗口

双击输出窗口左上角的图标，或单击输出窗口右上角的图标，都可关闭该输出窗口。如果窗口中的输出信息未存盘，系统显示提示对话框。输出信息存盘后，窗口关闭。

#### 5. 输出窗口能打开和保存的文件类型

输出窗口可以打开的文件类型有：Viewer document (\*.spv)输出文件、Syntax (\*.sps)SPSS 语句文件，Draft Viewer document (\*.rft)简化的输出文件，SPSS Script (\*.sbs)脚本文件，还有无格式的(\*.txt)文本文件。文本文件和其他各类型文件只能在窗口中编辑。

### 1.2.3 语句窗口

#### 1. 认识语句窗口

Syntax 语句窗口由以下 5 部分组成，如图 1-6 所示。

(1) 标题栏在窗口顶部，标有“SPSS Syntax Editor”。

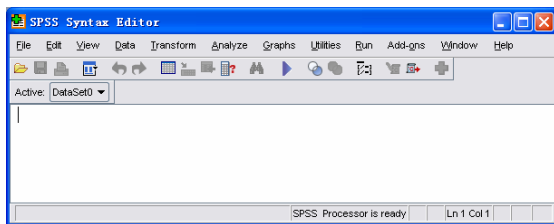


图 1-6 语句窗口

(2) 主菜单在标题栏下方，包括 File~Help 12 个菜单项。

(3) 功能图标按钮在主菜单下方，是可以简化操作的功能图标按钮，包括打开文件、保存文件、打印文件、调用最近使用过的对话框、删除或恢复上次操作、定位数据、定位观测、定位到变量、显示变量信息、查找、

运行当前命令、使用选择的变量集、显示语句帮助、生成/编辑草稿文件、运行草稿文件、主窗口标记（十字）等系统定制的图标按钮。

(4) 语句编辑区是图标下方的空白区域。在编辑区可以输入、编辑 SPSS 命令语句，构成 SPSS 程序，也可以输入和编辑文本文件。

(5) 状态行：语句窗口也有状态行，在窗口的最下面一行。

#### 2. 语句窗口的激活与功能

(1) 打开一个语句窗口的方法与步骤。

① 单击主菜单的 File 菜单项，展开下拉菜单。

② 单击下拉菜单中的 New 菜单项，在二级菜单中单击 Syntax 项，就打开了一个 Syntax 窗口，如图 1-6 所示。

(2) 建立语句窗口的另一种方法是当选择了一种统计分析方法，在相应的对话框、子对话框中设置程序参数后，在各可能生成命令程序的对话框中，单击 Paste 按钮，自动打开一个语句窗口，在语句窗口中生成与指定的统计分析方法及参数相应的 SPSS 命令语句。在语句窗口中可以对自动生成的命令语句进行编辑，熟悉 SPSS 语句的读者可

以增加对话框中不包括的参数或选项，然后提交系统执行。

### 3. 语句窗口的功能


(1) 各个 SPSS 过程的主对话框均有一个标有 **Paste** 的图标按钮。它把 SPSS 过程的命令语句，以及各选项对应的子命令语句，按照 SPSS 语言的语法组成一个或若干个完整的程序粘贴到主语句窗口中。

(2) 在语句窗口中可以使用键盘输入 SPSS 命令编写的 SPSS 程序，每个过程语句即一个完整的程序均以圆点“.”结束。


用 **Edit** 菜单项中的各种功能编辑窗口中的程序。

用 **File** 菜单项的各功能把窗口中的程序作为文件保存到磁盘中或关闭该窗口。

可以把已经存放在磁盘中的另一个程序文件调入，或独占该窗口或与已经存在于该窗口中的程序合并为一个程序作业，以便合并一次运行。

(3) 当使用鼠标选择一个完整的程序后，单击运行按钮，就把该窗口中选中的程序提交系统执行。

(4) 单击 **Syntax Help** 按钮，屏幕显示光标所在行上的命令或子命令所属的命令语句标准格式、可以选择的参数等，供读者查阅。

如果 **Syntax** 窗口中有多个过程语句，要执行其中的某一个过程，可以先用鼠标或键盘选择相应的语句，使之呈现反向显示，单击运行按钮即可提交系统执行。

### 4. 同时使用多个语句窗口

#### (1) 主语句窗口的概念

用前面介绍的从 **File** 菜单选择 **New** 的方法，可以同时打开若干语句窗口。在同一个 SPSS 期间首先打开并粘贴了语句的语句窗标为 **Syntax 1**，第二个打开并粘贴了语句的语句窗标为 **Syntax 2**……但只能有一个主窗口。主语句窗口标题栏为明亮或彩色的（默认为蓝色），十字号图标按钮为灰暗色。主语句窗口的功能有别于非主语句窗口，各过程对话框中所选的选择项形成的命令语句和子命令组成的程序只能粘贴到主语句窗口中。语句窗口只有被激活为主窗口，才可以在该窗口的编辑区输入程序语句，或将编辑区中的程序提交运行。

非激活窗口的标题底色为灰蓝色。十字号图标按钮为绿色。


#### (2) 将当前工作窗口变为主窗口

① 在一个 **Syntax** 窗口范围中单击鼠标左键，该窗口被激活成为当前工作语句窗口。

② 单击十字号图标，使其变为灰色。

③ 当屏幕画面上有两个以上语句窗口时，使用鼠标单击 **Utilities** 菜单，在下拉菜单中选择一个语句窗口，被选中的语句窗口变为主语句窗口。

#### (3) 关闭语句窗口

使用鼠标单击语句窗口左上角的 **SPSS** 语句窗口图标，选择下拉菜单中的“关闭”命令；使用复合控制键 **Alt+F4**；单击语句窗口右上角的图标，或选择 **File** 菜单项的



Exit 命令，在确定窗口中内容已经保存后，都可以关闭当前语句窗口。

5. 语句窗口打开与保存的文件类型是 SPSS 程序文件(\*.sps)。

### 1.2.4 窗口菜单

用鼠标单击 Window 菜单，展开一个下拉菜单，如图 1-7 所示。

#### 1. 选择窗口状态

单击 Window 菜单中 Minimize All Windows 命令，当前所有窗口最小化，即变成几个图标按钮显示在 Window 的状态栏内。

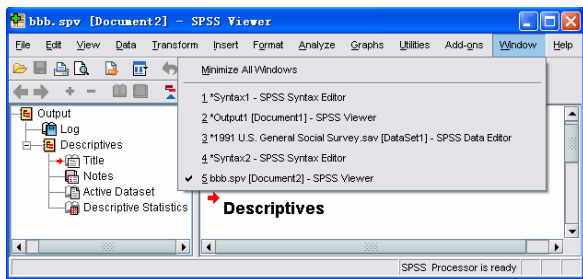


图 1-7 窗口菜单中的命令项

#### 2. 各窗口之间的切换

在 Minimize All Windows 命令下面是已经打开的窗口列表。打开的窗口名称前面，有对钩的窗口是主窗口，即当前工作窗口。没有对钩的窗口正处于非激活状态，图 1-7 中显示当前光标在输出窗中。

### 1.2.5 对话框及其使用方法

对话框，顾名思义就是提供人机对话环境和内容的窗口。主菜单中各项命令基本上是通过对话框中的选项、复选项、变量、参数、语句等操作来实现的，通过对话框中的各种功能按钮展开下拉菜单、执行命令或打开子对话框。

#### 1. 常见对话框类型

SPSS 中使用的对话框主要有如下三种。

##### (1) 文件操作对话框

例如打开已经存在的数据文件，按 File→Open→Data 顺序逐一单击鼠标左键，展开 Open File 对话框。与一般 Windows 应用软件的 Open File 对话框不同的是，SPSS 的打开文件对话框有 Paste 按钮，可以将打开文件的操作转换为命令语句粘贴到 Syntax 窗口中；而保存数据的窗口有选择要保存变量的功能。




##### (2) 统计分析主对话框

通过 Analyze 菜单中的各类统计分析命令所打开的第一个对话框，均为统计分析主对话框。在该对话框中选择参与分析的各类变量是该对话框的主要任务。另外分析方法不同会有不同的其他选项，例如选择分析中的算法以及输出选项等。图 1-8 为相关分析的主对话框。

SPSS 对话框中的变量表列出可以参与分析的变量标签，默认状态是变量名列在变量标签后面的中括号中。当变量标签与变量名太长，栏的宽度不够时，可以使用鼠标光标

指向该变量所在的行，该变量的变量标签和变量名则显示在该行的加长区中。

如果在系统参数设置对话框 **General** 选项卡的 **Variable List** 栏选择的是 **Name**，则在对话框变量表中只显示变量名。可以使用鼠标右键单击变量名，选择 **Variable Information** 项，查看变量标签。

尺度变量使用“”黄色尺子在左边做标记。分类变量的左端用“”三色条图标记。标称变量用“”三色彩球做标记，见图 1-8(a)。

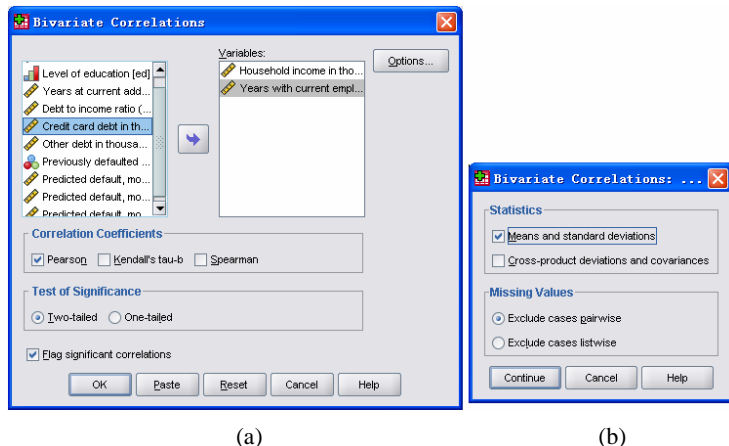


图 1-8 相关分析的主对话框和二级对话框

### (3) 其他选项对话框

其他选项对话框，即 **SPSS** 主菜单的其他菜单项对应的对话框或统计分析过程的二级对话框，这些对话框只在限定范围内提供选择的内容。图 1-8(b)所示的对话框是相关分析的二级对话框。

## 2. 对话框的构成

### (1) 按钮

按钮的主要功能是激活选项。它告诉系统去做什么，包括以下三类，见图 1-9。

① 移动变量按钮，见图 1-9(b)。按钮中央是箭头，它把变量表中选中的变量加到变量框中。例如选择参与分析的变量，指定分类变量，或者指定因变量、自变量等。该按钮有时也用在构成模型时的变量选择。按钮的指向是可以改变的。当使用鼠标键选择了原始变量（左面矩形框中）时，箭头按钮

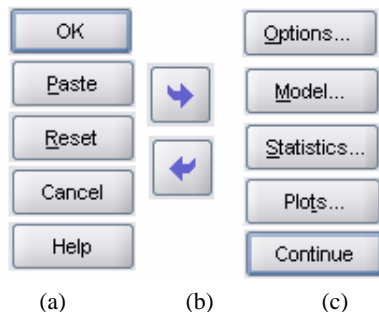


图 1-9 对话框中不同的功能按钮

指向右方，表示可以将选择的变量移到右边的变量表中去。当在右边的变量表中选择了变量时，箭头按钮指向左边，表示可以把变量表中的变量送回原始变量表中去，即从已

经选择参与分析的变量中剔除。

② 打开下一级对话框的按钮，见图 1-9 (c)中的 Options...按钮，其特点是按钮中字符后面有省略号，按钮中的单词是下一级对话框的名称。这类按钮常用的还有：Model 打开建立模型对话框的按钮；Plot 打开作图对话框的按钮；Statistics 打开统计量选择对话框的按钮和 Save 打开生成并保存新变量对话框的按钮等。

③ 执行功能按钮，每个对话框中都有这样几个执行功能按钮。

- OK 按钮，见图 1-9 (a)，单击这个按钮，把经过主菜单、子菜单、对话框，直到子对话框等选择的带有参数的命令过程语句提交系统执行。当选择或指定的变量、参数不符合运行相应过程的要求时，该按钮为灰色，不能提交系统运行。

- Paste 按钮，鼠标单击该按钮把通过对话框的各种操作组成的带有指定参数的过程命令语句显示到主语句窗口中。当选择或指定的变量、参数不符合运行相应过程的要求时，该按钮为灰色，表示没有具备生成可执行文件的条件。灰色按钮不能响应鼠标单击的操作。

- Reset 按钮，清除在对话框中进行的一切选择和设置，使其恢复到系统默认状态。

- Cancel 按钮，取消本次打开对话框后的操作，返回到上一级对话框或窗口。

- Help 按钮，打开帮助窗口，显示与当前对话框及其各项有关的帮助信息。

- Continue 按钮，一般是二级对话框中的按钮。单击该按钮表明确认在二级对话框中的参数选择，返回前一级对话框。与之并列的有 Cancel 按钮和 Help 按钮。

## (2) 选项

选项有两种。单选项形状像一个收音机旋钮，如图 1-10 所示。总是多个带有旋钮的选项排列在一起。这些选项只能择其一，不能同时选两个或两个以上。被选中的一项前面的圆圈旋钮中出现黑点。并列的若干项中必须选择其中的一项，而且只能选择一项。如果只有一项，无与之并列的项，选择与否均可。

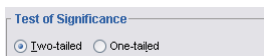


图 1-10 单选项

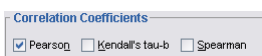


图 1-11 复选项

复选项形状为方框，被选中的复选项前有“√”出现，如图 1-11 所示。可以同时选中多个复选项，也可以一个不选。

任何一项都不选时，有时会产生不希望产生的结果，或者输出窗口中没有分析结果输出，甚至于出错。

## 1.2.6 设置工具栏中的工具图标按钮

各窗口中都有自己的工具栏，工具栏中显示常用功能的图标按钮，这些图标按钮使许多操作变得简单方便。

如果当前窗口中没有这些工具图标按钮，可以使用下述方法将这些工具图标按钮显示在各窗口工具栏中。下面以编辑数据窗中工具栏为例，将“复制”、“剪切”、“删除”

三个编辑工具添加到工具栏中。操作步骤与方法如下：

(1) 在数据编辑窗中，按 **View→Toolbars→Customize** 顺序逐一单击鼠标左键，展开 **Show Toolbars**（显示工具栏）对话框，如图 1-12 所示。

(2) 在 **Windows** 参数框内，单击向下箭头展开窗口表，由于每个窗口有不同的工具栏，要确定编辑哪个窗口的工具栏就在窗口下拉表中选择哪个窗口名。从 **Data Editor** 窗口的 **Viewer** 菜单启动 **Show Toolbars** 对话框，首先显示的是 **Data Editor**。

(3) 在 **Toolbars** 栏内显示的是在 **Window** 框中确定的窗口的工具栏名称。有的窗口可能同时出现两个以上工具栏名称选项。可以同时选择。但同时选择多个工具栏，会有重复的图标按钮出现在同一个窗口中。因此，最好使用系统默认的工具栏。

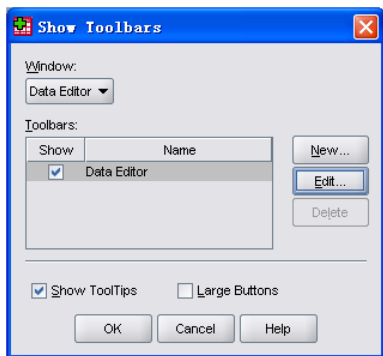


图 1-12 调用工具栏

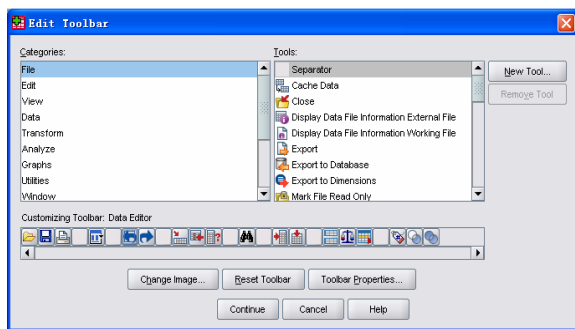


图 1-13 编辑工具栏对话框

(4) 在 **Toolbars** 栏内选择一个工具栏，使之显示彩色底纹。图 1-12 选择的是数据编辑窗口 **Data Editor**。

如果需要建立全新的工具栏，单击对话框右侧的 **New** 按钮。

(5) 单击右边的 **Edit...** 按钮，打开 **Edit Toolbar** 工具栏编辑对话框，见图 1-13。对话框分为三个部分，左面 **Categories** 工具分类栏列出的是当前窗口的菜单项。每一个菜单项对应的一组工具图标。当选择了一个菜单项时，所对应的所有工具图标显示在右面的 **Tools** 栏内。下面的 **Customizing Toolbar** 是当前窗口的工具栏，它包括了若干工具图标，是可以编辑的。

(6) 在 **Categories** 栏内选择一类工具，选择编辑工具，即选择 **Edit** 项，见图 1-14，在右边的 **Tools** 栏内显示编辑类的所有工具图标。

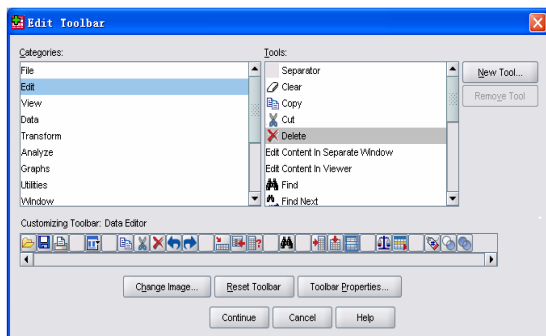







图 1-14 安排好的自定义工具栏

(7) 在对话框下边的 Customizing Toolbar: Data Editor, 工具栏为当前该工具栏的现状。可以用鼠标将某个图标拖曳到新的位置, 重新安排图标的排列。

(8) 在 Tools 栏内, 选择一个图标按钮。按下鼠标左键, 拖曳到下边的工具栏中, 松开鼠标键。选中的图标按钮出现在工具栏中。用这样的方法将图标:  Copy、 Cut、 Delete 一一拖曳到下面的工具栏中到 Undo 、Redo  图标按钮前边, 见图 1-14。

(9) 单击 Continue 按钮返回到如图 1-12 所示的 Show Toolbars 对话框中。单击 OK 按钮, 结束操作。此时数据窗中的工具栏已经增加了三个工具图标。

(10) 在 Edit Toolbar 窗口中编辑好的工具栏可以在其他窗口使用。单击对话框下面的 Toolbar Properties 展开对话框, 可以选择刚定义好的工具栏显示在哪个窗口。不熟悉窗口、工具栏等操作的读者对此功能慎重使用。如果需要重新安排, 单击该对话框下面的 Reset Toolbar (重置工具栏) 按钮, 恢复定义之前的工具栏状态。

## 1.3 系统参数设置

### 1.3.1 参数设置基本操作

系统初始状态和系统默认值的设置是通过 Edit→Options (参数设置) 对话框完成的, 通过 Edit 菜单中的 Options 命令打开该对话框。参数与状态设置生效的时间不同, 有的在确认后立即生效, 有的则要在下次启动 SPSS 系统时才生效。但无论何时生效, 只要生效, 设定的状态或参数即代替了原来系统给定的默认值。

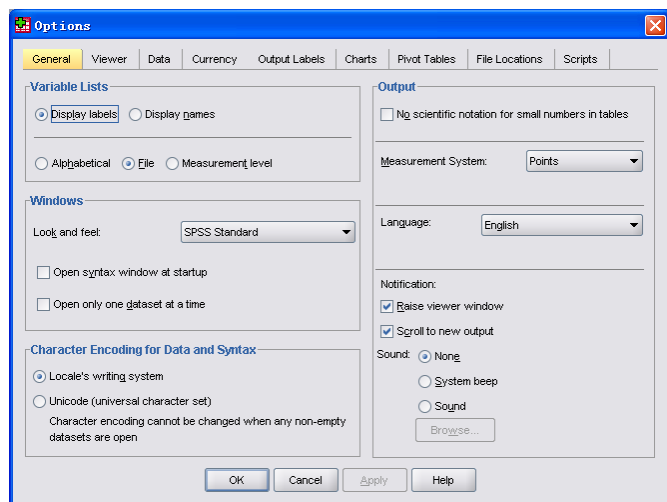


图 1-15 设置系统状态对话框及其通用参数选项卡

按 Edit→Options 顺序, 打开 Options 对话框, 在如图 1-15 所示的 Options 对话框中进行系统状态和参数的设置。有以下几种可能的情况需要使用对话框执行功能按钮。

(1) 当完成任何参数或状态设置后, 可单击 OK 按钮, 确认所作的设置并返回到 SPSS 主画面。

(2) 如果在 Options 对话框中一系列设置完成后认为设置得不够满意, 需要重新设置, 可以单击 Reset 按

钮恢复到打开该窗口时的原始设置状态，重新进行设置工作。

(3) 单击 **Cancel** 按钮退出 **Options** 对话框，返回到 **SPSS** 主画面，同时刚刚设置的参数作废。

(4) 单击 **Help** 按钮，展开与该对话框各项有关的帮助窗口，查看有关说明。

以上操作在每项设置过程中或完成后都可以进行，后续操作中类似设置不再重复。

### 1.3.2 通用参数设置

**General**（通用参数）选项卡上可设置各种通用参数，见图 1-15。

#### 1. 设置显示变量、顺序的方式

**General** 选项卡上的左边第一栏是标有 **Variable Lists** 栏，下面的单选项设定变量在变量表中的显示方式和显示顺序，可通过两组选项进行选择。

##### (1) 变量的显示方式

① **Display labels**，显示标签。选择此项，变量标签显示在前，变量名显示在后面的括号中。此为系统默认方式。

② **Display names**，显示变量名。选择此项，在各对话框的源变量表中只显示变量名。

##### (2) 变量的显示顺序

① **Alphabetical**，按变量名的字母顺序排列。

② **File**，按变量在数据文件中出现的顺序排列。此为系统默认方式。

改变变量显示顺序的设置对当前的工作数据文件无效，只对选择 **Apply** 和 **OK** 按钮以后打开的或定义的数据文件起作用。在各统计分析对话框中，源变量表中的变量，按选定的方式排列。

③ **Measurement level**，按变量的测度水平 **Nominal**、**Ordinal**、**Scale** 排列。

#### 2. 窗口状态的选择

**General** 选项卡左边第二栏，标有 **Windows**，在其中选择窗口状态。

(1) **Look and feel**：下拉菜单中有两项，可以选择其中之一：

① **SPSS Standard**，使用 **SPSS** 标准窗口。

② **Window**，一种具有表格线，颜色也不同的窗口。

(2) **Open syntax window at startup**：在启动 **SPSS** 时就打开语句窗口。习惯于使用 **SPSS** 语言编程和使用 **SPSS** 对话框功能的读者应该选择此项。

(3) **Open only one dataset at a time**：每次只打开一个数据集。选择此项，不能同时打开两个以上数据文件或数据窗口。

#### 3. Character Encoding for Data and Syntax，栏中数据和语句字符的选择

① **Locale's writing system**，选择此项使用当前的写作系统所用字符。

② **Unicode (universal character set)**，使用一种双字节的世界统一的编码系统的字符。当一个非空数据窗口在打开状态时，不能改变编码系统。

#### 4. Output 栏内的输出的设置

(1) No scientific notation for small numbers in tables, 在表格中对小数字不用科学计数法表示。

(2) Measurement System 参数框, 在下拉菜单中可以选择测度单位, 即 points (点或磅)、inch (英寸) 或 centimeters (厘米)。它们的换算关系为: 1 英寸=72 点=2.54 厘米。如果需要作精细的图形, 可以使用“点”作为单位, 系统默认单位为英寸。

(3) 在 Language 选项框中, 选择输出结果的默认语言, 常用的除英文外还有:

① Traditional Chinese, 输出使用繁体中文。如果没有安装繁体中文字库, 不要轻易设置, 否则结果会出现乱码。

② Simplified Chinese, 输出使用简体中文。输出表格标题或输出项有时所用术语翻译有误。

无论选择哪一个, 输出仍以英文为主, 只在输出表格标题和输出项使用指定的语言。

(4) Notification 栏控制在运行一个 SPSS 过程后在观察窗口中显示的输出结果的通告方式。有两个选项, 默认同时使用两种方式。

① Raise viewer window, 当有新处理结果时输出窗口自动弹出。

② Scroll to new output, 当有新处理结果时屏幕显示到新的输出信息处。

(5) Sound, 声音选项

① None, 产生输出信息时不发声。

② System beep, 产生输出信息时发出系统默认的声响, 以提醒读者。

③ Sound, 选择此项, 单击其下方的 Browse 按钮, 读者确定一个声音文件。产生输出时运行此文件, 发出声响以提醒读者。

### 1.3.3 输出观察窗口参数设置

在 Viewer 选项卡上设置观察窗即设置输出窗口的各种参数, 见图 1-16。在改变参数设置后, 单击 OK 按钮退出 Options 窗口后, 再次运行 SPSS 命令, 产生新的输出时才能生效。共有 4 部分的参数可根据需要重新设置。

#### 1. 初始输出状态设置

在 Viewer 选项卡的左边第一项, 标有 Initial Output State 栏, 在本栏中设置各种输出的初始状态。

(1) Item 参数框, 在该参数框中控制输出项在每次运行一个统计分析结果输出时, 是否自动显示或隐藏, 以及初始状态使用的对齐方式。可以选择的输出项有: Log (日志)、Warnings (警告信息)、Notes (注释信息)、Title (标题)、Page title (页标题)、Pivot tables (表格)、Charts (图表)、Text Output (文本输出信息) (表格中没有显示的输出信息)、Chart (图形)、Tree Model (树形结构图)。每选择一项, 就可以按下面(2)、(3)项设定该项的状态。

(2) 在 Contents are initially 下面的单选项: Shown (显示)、Hidden (隐藏), 确定 Item 框内所指定的项目是显示还是隐藏。

(3) 文本内容对齐方式。所有输出均默认左对齐, 仅打印输出的对齐方式由 Justification 下面的单选项确定: Align left (左对齐)、Center (居中对齐)、Align right (右对齐)。

(4) 选中最下面的 Display commands in the log, 在日志中显示 SPSS 命令, 读者可以从日志中复制命令语句并将它们保存在一个语句文件中。

## 2. 输出文本的字体、字号设置

在 Viewer 选项卡右面有三个栏目: Title 栏、Page Title 栏、Text Output 栏, 分别定义输出标题、页标题和输出文字的字体、字形、字号和颜色, 这些设置对新产生的输出生效。

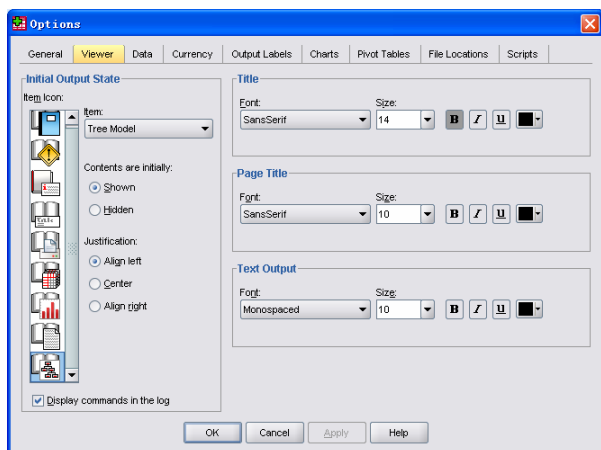


图 1-16 选择对话框中观察窗口选项卡

## 1.3.4 数据属性参数设置

Data 选项卡设置有关数据的各种参数, 如图 1-17 所示。

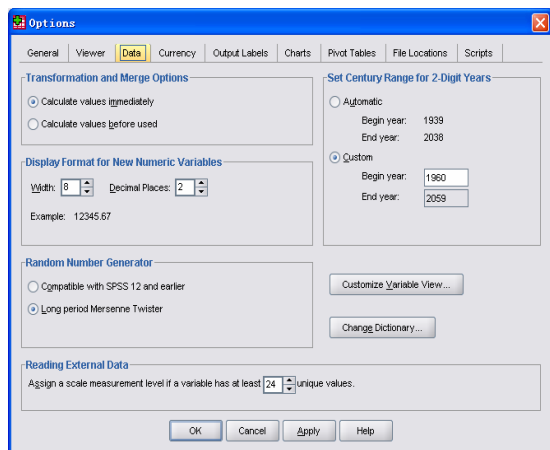


图 1-17 选择对话框中的数据选项卡

### 1. Transformation and Merge Options 栏, 选择数据转换和合并

SPSS 进行某些数据转换 (如 Compute 和 Recode) 和文件转换 (如增加变量或观测量) 后, 可以不要求立即执行, 而是到 SPSS 读取这些数据去执行另一个命令时再对数据或文件进行转换。何时执行转换, 可以通过下面的选项确定。

(1) Calculate values immediately, 要求指定转换方法后立即执行。

(2) Calculate values before used, 指定在使用之前再进行转换或合并。

对一个大的数据文件, 选择此项可以延迟执行以便节省处理时间。

2. Display Format for New Numeric Variables 栏, 即为新数值变量指定系统默认的显



## 示宽度和小数位数

如果一个数值相对于显示格式太长，SPSS 首先截掉小数部分，然后转化成科学记数法显示。显示格式对参与计算的数值本身没有影响，例如 123456.78 可以显示成 123456，但在进行任何计算时都使用未被截掉小数部分的原始值。

(1) 在 Width 参数框中输入显示数值的总宽度。

(2) 在 Decimal Places 参数框中输入显示数值的小数位数。

### 3. Random Number Generator 随机数生成器设置

(1) Compatible with SPSS 12 and earlier 与 SPSS 12 和以前的版本兼容的随机数发生器。如果需要使用 12 版以前版本的随机数发生器产生使用指定种子数的随机数，就选择此项。

(2) Long period Mersenne Twister 更可靠的新随机数发生器。

4. Set Century Range for 2-Digit Years 栏，即对日期型数据中的年份指定使用两位数字输入和显示（例如：10/28/97、29-OCT-96）

(1) Automatic，自动指定表示年限范围项，根据当前年向前 69 年作为开始，向后 30 年作为结束。当前年即系统时间确定的年，加上当前的一年共 100 年的范围。

(2) Custom，自定义范围项。在 Begin year 参数框中读者可以输入范围的起始年。End year 参数框中的数值是系统自动确定的，当输入了起始年后，单击 Apply 按钮，自动计算并显示结束年。如图 1-17 就是输入起始年 1960，结束年显示为 2059 年，范围也是 100 年。

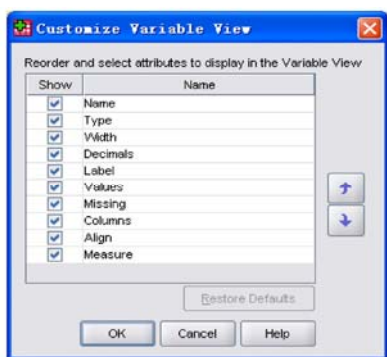


图 1-18 自定义变量默认属性对话框

拉列表中选择一种语言，该语言字典将用于数据窗中检查拼写。但是无中文字典可以选择。

7. Reading External Data，在该栏中定义尺度数据与分类数据的界限。如果默认一个数值型变量至少有 24 个不同的数值，则认为它是 Scale 尺度变量。可以

### 5. Customize Variable View...按钮

单击该按钮，展开自定义变量观察窗的对话框，如图 1-18 所示。在该对话框中重新安排和选择变量观察窗中的变量属性项。

(1) 选择一项，单击向上或向下箭头按钮，可以改变所选项的位置。

(2) 单击某项前的方框，有对钩，该项显示在数据观察窗中，没有对钩则不显示。

6. Change Dictionary...，单击该按钮打开自定义字典对话框。

见图 1-19。在下

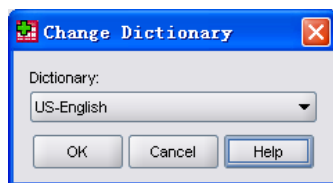


图 1-19 定义字典对话框

单击上下箭头按钮增加或减少这个数值，更改这个参数。

### 1.3.5 货币变量自定义格式设置

SPSS 允许读者自己设定常用的货币数值型变量的格式。Currency 选项卡设置有关数据的各种参数，如图 1-20 所示。

#### 1. Custom Output Formats 栏，读者定义输出格式

在该栏中列出可以设置的五种自定义格式，分别命名为 CCA、CCB、CCC、CCD 和 CCE。选择一个名称，例如图中选择 CCC。再做下面的操作：

#### 2. All Values 栏，设置数值的首尾字符

(1) Prefix 前缀框，设置在数值的前面添加的字符，系统默认值是空格。我们设置“\$”。

(2) Suffix 后缀框，设置在数值的后面添加的字符，系统默认值是空格。

#### 3. Negative Values 栏，设置负数的首尾字符

(1) Prefix 前缀框。在该框内，输入负数首字符，系统默认值是“-”。

(2) Suffix 后缀框。在该框内，输入尾字符，默认值是空格。

#### 4. Decimal Separator，设置十进制数的小数点符号

(1) Period，用圆点作为小数点符号，每三位分隔符为逗号。此为系统默认值。

(2) Comma，用逗号作小数点符号，每三位分隔符为圆点。

以上参数设置完毕，按格式表达的数字样例显示在 Sample 矩形框中。在图 1-20 中，上面一个是正数的样例，下面一个是负数的样例。选择格式的命名后即可按 Apply 按钮确认定义的格式。定义的格式即可在定义数值型变量时使用。

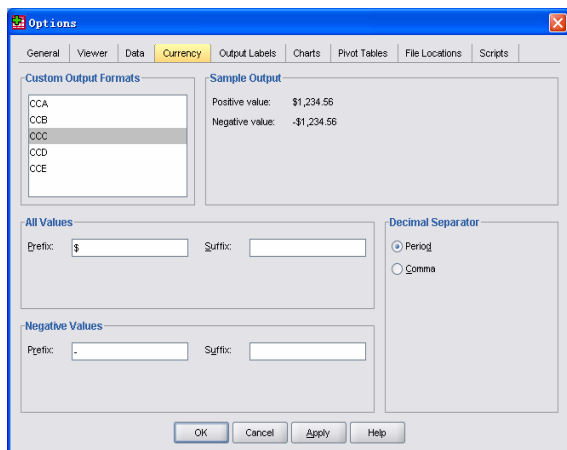


图 1-20 数值型变量自定义格式选项卡

### 1.3.6 标签输出设置

使用输出结果或表格格式，将变量标签或变量值等一并输出，可以让读者很方便地阅读这些结果和表格。这些变量标签或变量值标签都是在定义一个变量时，使用主菜单 Data 中的 Define Variables 功能项定义的。Output Labels 选项卡设置标签输出的各种参数，如图 1-21 所示。输出的表格是显示变量名还是显示变量标签，遇到需要显示分类变量值时是显示变量值还是显示它的值标签，可以使用 Output Labels 选项卡中的两个栏目来设定。

1. 在 Outline Labeling 结果标签栏中，设定输出表格时在相应的导航栏中是否使用标签

(1) Variables in item labels shown as 选项栏，设置变量标识。在下拉列表中有 3 个选项，指定其中一个。

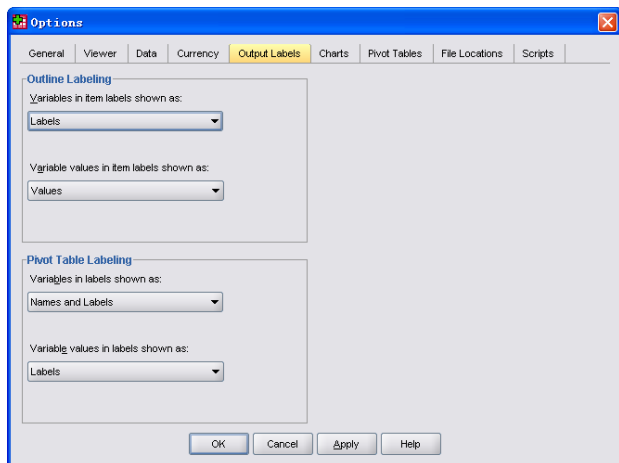


图 1-21 选择对话框中标签输出选项卡

① Labels，选择此项使用变量标签标识每个变量。

② Name，选择此项使用变量名标识每个变量。

③ Names and Labels，选择此项同时使用变量名和变量标签标识一个变量。

(2) 在 Variable values in item labels shown as 栏，设置变量值标识。单击向下箭头按钮，在下拉列表中有 3 个选项，指定其中一个。

① Labels，选择此项使用变量值标签标识每个变量值。

② Values，选择此项使用变量

值本身标识每个变量值。

③ Values and Labels，选择此项同时使用变量值和变量值标签标识每个变量值。

2. 在 Pivot Table Labeling 栏，设置表格标签

设定输出表格时是否使用标签。其操作过程与 Outline Labeling 栏相同。

**注意：**当变量标签或值标签过长时，在图形或表格中使用标签不一定是合适的。因此使用标签与否，要视实际情况而定。

输出标签选项只有对指定这些选项之后产生的输出生效。对当前已经在输出窗口中的输出图形或表格不起作用。

### 1.3.7 统计图形参数设置

Charts 选项卡设置统计图形的各种参数，如图 1-22 所示。

1. Chart Template 栏，设置图形模板，新图形可以套用这些参数

(1) Use current settings，使用当前系统默认的模板和此选项卡中的默认参数。

(2) Use chart template file，使用图形模板文件中的参数。选择此项，需要单击 Browse 按钮，在“打开”对话框中指定一个模板文件。

也可以建立新的模板文件，用需要的参数生成图形并将其保存到模板文件中。方法是生成图形后，双击图形，打开 Chart Editor 对话框。在 File 菜单中选择 Save Chart

Template 命令, 设定保存的模板项目后把图形保存为扩展名为 “.sgt” 的模板文件。

## 2. Chart Aspect Ratio 栏, 设置图形的宽高比

默认值为 1.25。在该参数框中可以直接输入需要的比例数值。输入的数值在 0.1~10.0 之间, 设置比例小于 1,

图形高度大于图的宽度; 比例大于 1, 图宽大于图高; 比例等于 1, 图形为高宽相等的正方形。一旦图形生成, 在 SPSS 中其长宽比就不能改变了。

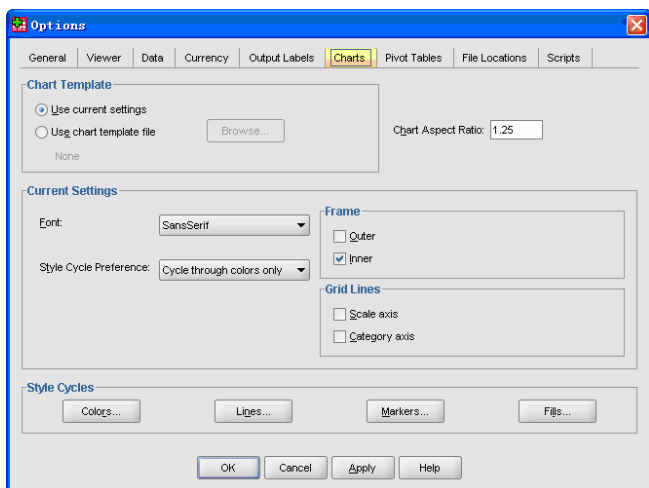


图 1-22 统计图形参数设置选项卡

## 3. Current Settings 栏, 设置当前生成图形参数

(1) Font 栏, 设置图形中的文字字体, 单击向下箭头, 在下拉列表中选择一种字体。默认的是没有修饰的普通字体。

(2) Style Cycle Preference 选项框, 在其中设置新生成图形的填充方式。

- Cycle through colors, then patterns, 用不同颜色区别图形不同的分类, 不使用底纹图案。

- Cycle through patterns only, 只用不同底纹区别图形不同的分类, 而不使用颜色。如果显示器为单色显示器, 选择此项可以在屏幕上获得比较好的图形显示效果。

(3) Frame 图形框设置栏, 本栏内有两个复选项。

- Outer 外框, 选中后在整个统计图 (包含标题和图例等) 的外围加框。

- Inner 内框, 选中后只对统计图形加边框。

(4) Grid Lines 格线栏, 提供两种坐标轴格线。

- Scale axis 刻度轴格线, 选中后显示刻度坐标轴格线。

- Category axis 分类轴格线, 选中后显示分类坐标轴格线。

## 4. Style Cycles 栏, 改变图形外观样式

按下 Colors、Lines、Markers 和 Fills 按钮打开相应的对话框, 设置图形的颜色、线条、标记和填充的样式。

### (1) 设置图形颜色

单击 Colors 按钮, 打开如图 1-23 的 Data Element Colors 数据元素颜色对话框。左边是 Styles to Edit 装饰编辑栏。

① Simple Charts 是默认的, 简单图形用单一颜色标注所有图形元素。默认颜色是淡

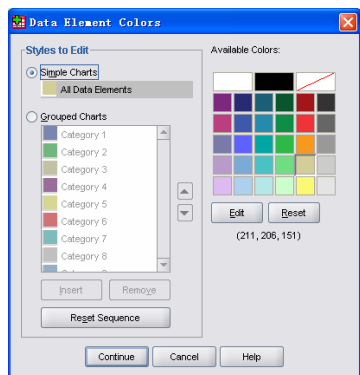


图 1-23 数据元素颜色设置对话框

黄色显示在该项目下方。例如做出的柱形图每个柱都是淡黄色。要改变默认颜色，只要在右面的调色板中选中一种合适的颜色，双击之即可。

② **Grouped Charts**, 对于一组图形需要有不同颜色表示时，可选择此项。图形颜色的选取可按下面栏中自上而下的顺序选择。

- 如果要改变颜色使用顺序，只要单击选择的一种颜色，然后单击向上或向下箭头按钮，就可以向前、向后移动使用顺序。
- 如果改变栏中的某一种颜色，只要单击这个颜色块，然后到右面的调色板中选择一种合适的颜色单击之即可。

• 可以在 **Available Colors** 中选择一种颜色，单击 **Insert** 按钮，将其插入到颜色列表中；对不想使用的颜色，在列表中选后单击 **Remove** 按钮，将其放回 **Available Colors** 可用颜色列表中。

③ 编辑调色板有三种方式。单击调色板下方的 **Edit** 按钮，打开如图 1-24(a)所示的 **Choose a color** 对话框，三个选项卡是选择合适颜色的三种方式。

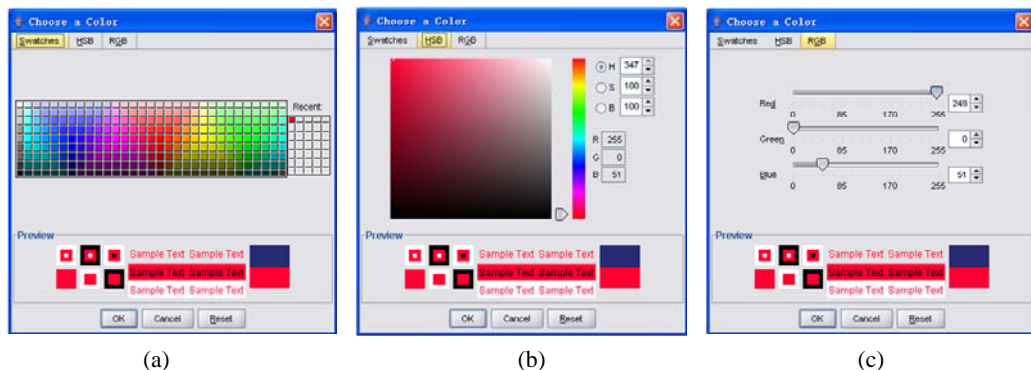


图 1-24 三种方式改变调色板颜色的对话框

• 第一种方式是通过 **Swatches** 样品块方式选择颜色，只要鼠标单击选中的颜色，就有一小块样品显示在右侧的格子中。在下面的 **Preview** 栏中察看各种符号颜色是否合适。

• 第二种方式是 **HSB** 色相、饱和度、明度方式。方块中是在样品快方式中选择的颜色，只不过饱和度和明度是渐变的。只要用鼠标拖曳其中的圆圈到饱和度和明度合适位置松开鼠标即可。是否合适，察看下面的预览窗口中的各种符号。

- 第三种方式是 RGB 方式即红绿蓝基本色参数方式。熟悉参数的读者可以使用这种方式。

一般凭直觉调整颜色的读者只要选用前两种方式就可以满足要求。设置颜色完成后单击 OK 按钮返回 Data Element Colors 数据元素颜色对话框。

单击 Continue 按钮返回主对话框。

## (2) 设置线型

单击 Lines 按钮打开 Data Element Lines 数据元素线型对话框, 如图 1-25 所示。

在 Style to Edit 栏内有两种目标供选择:

① Simple Charts (简单图形)。使用的默认线型是细直线。样品显示在选择项下方。若要改变这个基本线型, 只要在右面的线型列表中选择一种, 单击之即可。例如只有一根折线的折线图, 可以选择这种方式。

② Grouped Charts (成组图形)。例如由两条以上的直线或折线组成的图形, 需要两种以上线型, 以便区别, 则可以选择此项。

- 改变线型使用顺序, 线型使用顺序按 Grouped Charts 下的线型列表自上而下的顺序。要想改变顺序只要选择一种线型单击之, 然后单击向上或向下箭头按钮移动该线型。

- 也可以用 Available Lines 中选中的线型提供单击 Insert 按钮插入到列表中所选线型的下方。对不要使用的线型, 可以在列表中选择后, 单击 Remove 按钮将其送回 Available Lines 中。

- 改变列表中某线型, 只要单击该线型, 然后在右面 Available Lines 可选择的线型中选好一种单击之即可。

- 若要恢复默认状态只需要单击 Reset Sequence 按钮。

## (3) 设置标记

单击图 1-22 中的 Markers 按钮, 打开 Data Element Markers 设置数据标记对话框, 见图 1-26(a)。设置图形中数据点所用标记。

## (4) 设置图案

单击图 1-22 中的 Fills 按钮, 打开 Data Element Fills 对话框, 见图 1-26(b)为设置图案对话框。设置有框图形, 如柱形、饼图等, 填充内部使用的图案或称底纹。操作方法与设置线型相同。

以上所有在 Chart 选项卡中的设置在单击 OK 按钮后, 只对设置后产生的图形生效。

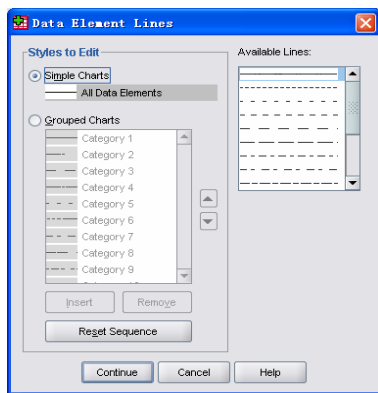


图 1-25 设置数据元素线型对话框



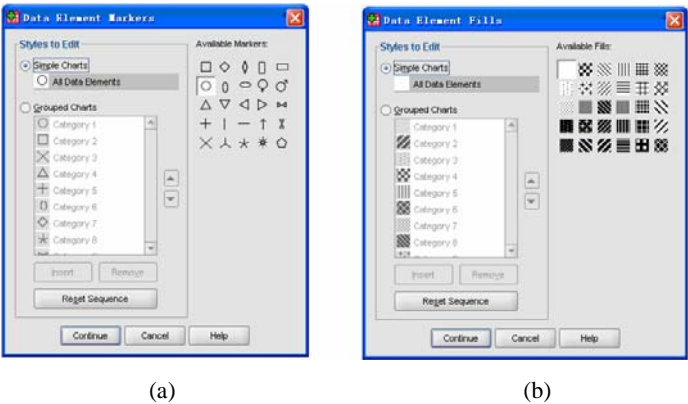


图 1-26 设置标记和图案的对话框

1.3.8 输出表格参数设置

Pivot Tables 选项卡，设置默认表格样式及有关参数，如图 1-27 所示。在表格参数

设置选项卡中，对新的表格输出设置外观，其中 Table Look 栏可以设置各种表格属性，包括显示器和网格线宽度、字体、字号和颜色、背景颜色等。

(1) Table Look 栏。在该栏中选择一个表格外观样式，被选择的表格样式显示在 Sample 下的矩形框中。单击 Apply 或 OK 按钮，新表格按选择的形式生成。Table Look 中的表格样式文件保存在 SPSS 系统所安装的 Look 文件夹中。选中的表格样式文件的

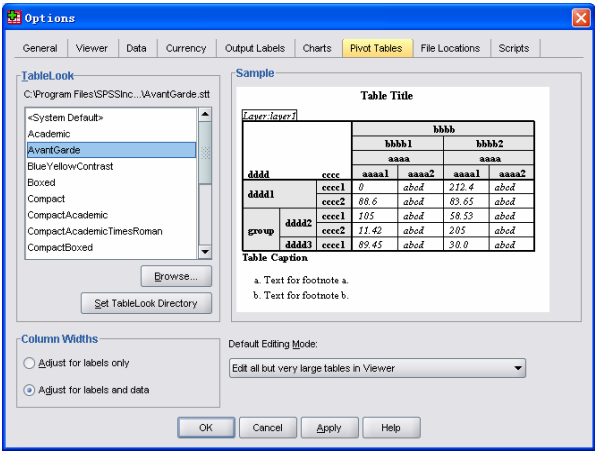


图 1-27 选择对话框中表格选项卡

保存位置和文件名，显示在 Table Look 栏目中第一行上。默认的是外框粗实线。内部细实线，不加任何修饰的普通表格。

我们还可以使用 SPSS 提供的 Table Looks 功能建立表格样式，即双击输出窗口中的表格，在表格编辑窗口中选择 Format 菜单的 Table Looks 命令，展开 Table Looks 对话框，选择一种基本样式。再利用 Edit Look 按钮展开 Table Properties 对话框，改变表格各部位的参数建立自己的表格样式。

(2) 单击 Browse 按钮展开相应的对话框，选择搜索路径，指定扩展名为\*.stt 的表格样式文件。

(3) 单击 Set Table Look Directory 按钮, 可以改变系统默认的 Table Look 文件夹。安装 SPSS 系统时, 默认安装在 C:\Program files\SPSSInc\spss16\Look 文件夹中, 如果没有按此文件夹安装, 应该改变路径, 以便显示系统给出的表格样式文件。单击 Set Table Look Directory 按钮, 在展开的对话框中指定文件夹。

(4) Column Widths 栏设置表格列宽度两个单项选项:

① Adjust for labels only, 调整列宽到该标签的宽度。这种调整方式可以产生比较紧凑的表格。当数值宽度大于标签宽度时就不会显示, 并用星号表明要显示的值太宽。

② Adjust for labels and data, 即标签与宽度最大的值中哪个值所占宽度更宽, 列宽则自动调整到这个宽度。这个选项将生成较宽的表格, 但能保证所有的标签和变量值均能显示出来。

(5) Default Editing Mode 栏, 可从下拉列表中选择设置默认的编辑表格模式。

① Edit all but very very large tables in Viewer, 在输出观察窗中除很大的表格外, 编辑所有表格。

② Edit all tables in Viewer, 在输出观察窗中编辑所有表格。

③ Open all tables in a separate window, 在不同窗口中打开所有表格。

单击 OK 按钮, 退出 Options 对话框, 所设置的表格样式只对以后产生的表格生效。

### 1.3.9 文件默认存取位置设置

File Locations 选项卡中的设置项控制 SPSS 启动后打开和保存文件的位置。系统默认的位置显示在每个选择项的编辑栏中。一般都是作为 Windows 用户的 My Document 文件夹。可以单击 Browse 按钮对文件位置进行重新设置。此功能为减少启动 SPSS 后查找数据文件或其他类型文件的操作而设置。

File Locations 选项卡如图 1-28 所示。

(1) Startup Folders for Open and Save dialogs, 该栏中设定打开和保存对话框所启动的默认文件夹。

**注意:** 指定的文件夹必须事先建好。如果指定了一个不存在的文件夹, 单击 Set 按钮后, 系统会给出警告信

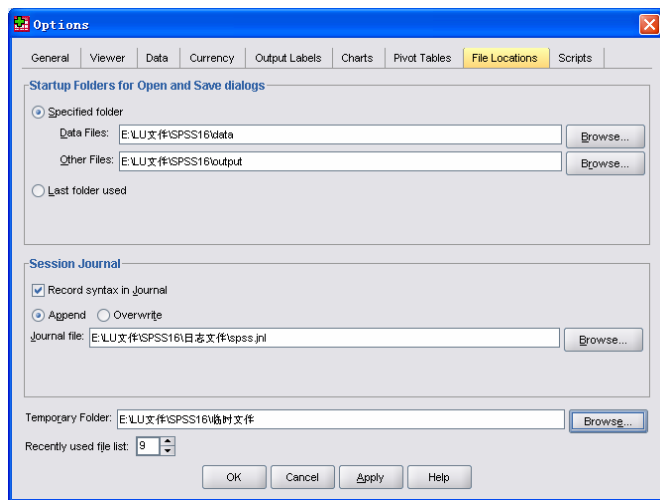


图 1-28 设置文件打开、保存位置的选项卡



息，并要求改变。

① Specified folder，在该栏中指定文件夹：

- 在 Data Files 栏直接输入或通过 Brows 定位，指定数据文件位置。即单击 Browse 按钮，打开 Default Data Folder 对话框，改变系统默认的设置。在 look in 下拉列表中确定文件夹位置，在 Folder name 框中输入文件夹的名字，单击 Set 按钮完成设置。

- 在 Other Files 栏直接输入或通过 Brows 定位，指定非数据文件位置。

② Last folder used，选择此项，启动 SPSS 后，打开或保存操作直接在打开或保存对话框中自动定位的文件夹，即上次从 SPSS 退出时最后使用的文件夹。

(2) Session Journal，指定 SPSS 运行时产生的日志文件自动保存的位置和形式。

① Record Syntax in Journal，选择此项，每次运行会把语句写进日志文件。系统默认选择此项，习惯于编程的人员更要选择此项。日志中记录的语句包括写在 Syntax 窗中的程序，也包括对话框操作时调用的命令、设置的参数等形成的语句程序。对下一次修改程序减少程序输入量很有用。

② 日志文件续写方式设定：

- Append，每次运行的语句接在前一次运行语句记录后面，存入日志文件。

- Overwrite，每次运行语句存入日志文件时覆盖前一次存入的内容。

③ 设定日志文件名及存储路径。在 Journal file 后面直接输入或单击 Browse 按钮，展开保存日志文件的 Save as 对话框，指定保存日志文件的存储位置和文件名。

(3) Temporary Folder（临时文件路径设置）。读者可以直接输入或者单击 Browse 按钮，打开 Temporary Folder 对话框，设置在统计处理过程中的临时文件的存放位置和文件名。临时文件往往需要较大空间，例如 200MB 的数据文件，需要大于 400MB 的临时文件空间。不用的临时文件及时删除。

(4) Recently used file list（最近使用过的文件数设定）。该栏设定最近使用文件数目。它控制显示在文件菜单中的文件名数目。改变 Entries 参数框中的数字，即可达到目的。

## 1.4 统计分析功能概述

SPSS 16.0 的统计分析功能主要集中在以下三个方面。

### 1. 统计分析函数

统计分析函数共 18 类 195 个函数。如算术函数、CDF 与非中心 CDF 函数、转换函数、当前日期时间函数、日期算法函数、日期生成函数、日期提取函数、反 DF 函数、PDF 与非中心 PDF 函数、随机数函数、查找函数、显著性函数、统计函数、字符函数、时间间隔生成函数、时间间隔提取函数和缺失值函数，等等。

## 2. 统计分析过程

在 **Analysis** 菜单中有 22 类, 73 个分析过程。另外还有可以使用语句实现的分析而没有收入窗口化的 **SPSS** 软件中的统计方法。在窗口化软件中的方法都可以使用编程语句实现。

## 3. 统计图

统计图可以直观表达数据特征和统计分析的结果。大致可以分为以下两类。

(1) 在 **Graph** 菜单中包括条形图、线图、面积图、圆图、高低图、帕累托图、控制图、箱图、误差条图、散点图、直方图、**P-P** 概率图、**Q-Q** 概率图、序列图等, 并有一套灵活、完整的对统计图进行编辑的方法。这些统计图是对数据统计特征的描述, 可以作为初步的统计分析和对数据特征的认识。

(2) 绝大多数统计分析方法都能产生统计图。这些图有些可以通过 **Graph** 菜单中的图形功能产生, 有的则直观表达统计分析的结果, 一般都在分析过程对话框的二级窗口 **Plot** 中的选项中。

注意各种统计分析方法使用的条件, 正确选择和充分利用 **SPSS** 中的各种统计分析功能, 辛辛苦苦获得的数据通过定量分析, 一定能挖掘出有用的信息。

# 1.5 数据与变量

## 1.5.1 常量与变量

### 1. SPSS 常量

常用的 **SPSS** 常量有数值型、字符型、日期型和日期时间型。

(1) 数值型常量就是 **SPSS** 语句中的数字。一般使用两种书写方式, 一种是普通书写方式, 例如 26、38.4 等。另一种是科学记数法, 即用指数表示数值的计算机书写方式, 用于表示特别大或特别小的数字, 例如 1.23E18 (或 1.23E+18) 表示  $1.23 \times 10^{18}$ , 2.56E-16 表示  $2.56 \times 10^{-16}$ 。

(2) 字符串常量是被单引号或双引号括起来的一串字符。如果字符串中带有“'”字符, 则该字符串常量必须使用双引号括起来, 例如“BOY'S BOOK”。在数据窗口中的字符串不使用引号。

(3) 日期型常量表示方法很多, 可以使用表 1-1 中所列的各种格式。

### 2. SPSS 变量及其属性

**SPSS** 中的变量除应定义变量名外还应该定义 4 个属性: 变量类型 (type); 格式——变量宽度 (width)、小数位数 (decimal); 缺失值定义 (missing value); 测度类型 (measure)。另外为输出查看方便还可以定义变量标签 (label) 和值标签 (values); 变量在数据窗口中的显示宽度 (columns)、对齐方式 (Align); **SPSS** 的变量至少要定义变量名和变量类

型, 其他属性可以采用默认值。

(1) 变量名命名应该遵循的原则

① SPSS 的变量名最多可长达 64 个字节, 相当于 64 个英文字符或 32 个汉字的长度。

② 首字符不能是数字, 必须字母打头, 其后可为除 “?”、“-”、“!” “\*”、“#”、“\$” 和空格以外的字符或数字。但应该注意, 不能以下画线 “\_” 和圆点 “.” 作为自定义变量名的最后一个字符。

③ 变量名不能与 SPSS 保留字相同。SPSS 的保留字包括: ALL、AND、BY、EQ、GE、GT、LE、LT、NE、NOT、OR、TO、WITH。

④ 系统不区分变量名中的大小写字符, 例如 ABC 和 abc 被认为是同一个变量。

(2) 变量类型与默认长度

SPSS 变量有三种基本类型: 数值型、字符型、日期型 (或日期时间型)。数值型变量又按不同要求分为五种。系统默认的变量类型为标准数值型变量 (Numeric)。每种类型的变量由系统给定默认长度。小数点或其他分界符包括在总长度之内。变量的系统默认长度可以用 Edit 菜单中的 Options 命令重新设置。

① 标准数值型变量 (Numeric), 默认总长度为 8, 小数位数为 2。标准数值型变量的值用标准数值格式显示, 小数点用圆点, 可以用标准数值格式输入, 也可以用科学记数法输入。使用科学记数法输入时, 显示出来的还是标准数值格式的数值。

② 带逗点的数值型变量 (Comma), 默认总长度为 8, 小数位数为 2。其值在显示时整数部分自右向左每三位用一个逗点作分隔符, 用圆点作小数点。定义为此格式的变量, 在输入时可以带逗点, 也可以不带逗点, 还可以用科学记数法输入。使用科学记数法输入时, 显示的还是用圆点作小数点, 逗点作三位分隔符的数值。

③ 圆点数值型变量 (Dot), 默认总长度为 8, 小数位数为 2。显示方式与带逗点的数值型变量正好相反。整数部分自右向左每三位用一个圆点作分隔符, 用逗点作小数与整数间的分界符。输入时可以带圆点, 也可以不带圆点。还可以用科学记数法输入。

④ 科学记数法 (Scientific Notation), 默认总长度为 8, 小数位数为 2。

对于数值很大或很小的变量可以使用科学记数法, 这种变量的值可以有指数部分也可以没有指数部分。表示指数的字母可以用 E, 也可以用 D。指数部分可以带正负号, 正号可以省略, 甚至指数部分不用字母 D 或 E, 只用符号表示也是可以接受的。例如表示一百二十三, 可以用以下方式输入: 1.23E2、123、1.23D2、1.23E+2、1.23+2 等。

⑤ 带美元符号的数值型变量 (Dollar), 默认总长度为 8, 小数位数为 2。其值在显示时有效数字前带有 “\$”, 变量总长度包括 “\$” 符号在内, 其余规定与标准数值格式相同。输入数据时可以带 “\$”, 也可以不带。显示在数据表格中的数值由系统自动加上 “\$” 符号和分隔符。可以用科学记数法输入, 如果数值不超过定义的长度, 则显示在数据表格中的数值自动变换为定义的格式。

带美元符号的数值型变量的具体格式还可以从格式列表框中选择，见表 1-1。

表 1-1 带美元符号的数值型变量格式列表框选项举例

格 式	总长度	小数位数	格 式	总长度	小数位数
\$##	3	0	\$#####	7	2
\$#,###	6	0	\$#,###.##	9	2

选定的格式只对在数据表格中的显示形式有效，当输入的数值小数位数超过格式规定时，系统自动进行四舍五入处理。如果输入的整数位数超出规定的格式，显示时自动去掉作为三位分隔符的逗号。

读者应该根据数据中最大数值的位数指定显示格式，以便使显示与输入的值一致。

⑥ 自定义型 (Custom Currency) 是一种由读者利用 Edit 菜单的 Options 功能来定义的，定义方法参见 1.3 节。

⑦ 日期型变量 (Date)

SPSS 的日期型变量可以表示日期，也可以表示时间。日期型变量的值按指定的格式输入和显示，不能直接参与运算。要想使用日期型变量的值进行运算，必须通过有关的日期函数转换。详见第 5 章。

(3) 变量格式

对数据的宽度 width 和小数位数 decimal 的要求。对数值型变量要定义宽度和小数位数。对字符型变量只定义宽度。日期型一般使用默认宽度，一旦日期格式确定了，宽度就确定了，不用再进行设置。

(4) 变量的标签与值标签

① 变量标签 (Variable Labels) 是对变量名附加的进一步说明。变量名只能由不超过 64 个字符组成，如果 64 个字符不足以表明变量的含义，或变量比较多时，则需要用变量标签对变量名的含义加以解释。如果 SPSS 运行在中文环境下，也可以给变量名附加中文标签，见表 1-2。

表 1-2 变量和变量值标签

变 量	变 量 标 签	变 量 值	变量值标签
Gender	性别	f	男
		m	女
Height	身高	1	<=1.49m
		2	1.50~1.59m
		3	1.60~1.69m
		4	1.70~1.79m
		5	>=1.80m

② 变量值标签 (Values) 是对变量可能取值附加的进一步说明。对分类变量往往要

定义其取值的标签。如果 SPSS 运行在简体中文版的 Windows 环境下，也可以给变量值附加中文标签，见表 1-2。

变量标签和变量值标签是可选择的属性，可定义，也可不定义。为了对输出信息进行解释并得出结论，建议使用中文标签。在输出窗口的输出表格中可以使用标签表明变量和变量值，这就要在 Edit 菜单中的 Options 功能对话框中进行设置。

### (5) 变量的显示格式

① 宽度 (columns) 显示数据的宽度。应该区分定义变量类型时指定的宽度与定义显示格式时的宽度。显示宽度应该综合考虑变量类型定义的总长度和变量名所占宽度。显示宽度不影响机内值，不影响分析运算结果，只影响显示。

② 对齐方式：分为左对齐、右对齐、中间对齐。一般情况下，数值型变量默认右对齐；字符型变量默认左对齐，也可以指定为中间对齐方式。

### (6) 缺失值 (Missing)。

已经输入的失真数据、没有测到或没有记录的数据，以特殊的数字或符号输入到数据文件中，统称为“缺失值”，都应该加以定义。在分析时不能使用，或需要单独处理。在 SPSS 中，字符型变量默认的缺失值为空格；数值型变量的缺失值没有默认值，需要定义。各分析过程对缺失值的处理都有默认的方法，也可以由读者指定选择项，定义如何处理这些缺失值。

### (7) 变量测度方式 Measure 是指变量是如何测量的

① 等间隔测度变量，即按与尺度的比例测度的变量，也可称为 Scale 尺度变量，如身高、体重。

② 有序变量 Ordinal，如表示职称、职务、对某事物的赞同程度的变量，是分类变量中有顺序特性的一种。可以用有序的数字作为代码。设置了值标签的变量被认为是有序的分类变量。可以作为分组变量，也可以参与某些分析过程的运算。

③ 名义变量 Nominal，是无序的分类变量，取值是无法度量的。只能作为分组变量使用。如表示民族、宗教信仰、党派等的变量。

分类变量值为数值时，它与尺度变量的分界默认值为 24。当变量的独立数值的个数大于 24 时被认为是尺度变量，小于 24 时认为是有序的分类变量。这个数值也可以在 Edit 菜单中的 Options 中重新设定。

## 1.5.2 操作符与表达式

SPSS 的基本运算共有三种：数学运算、关系运算、逻辑运算。运算符见表 1-3。

### 1. 算术运算符与算术表达式

算术运算符可以连接数值型的常量、变量和函数构成算术表达式。其运算结果为数值型常量。例如  $X + Y * 2 / (A + B) - 1 + \text{ABS}(A * Z)$  就是一个合法的算术表达式。在算术运算表

达式中,运算的优先顺序为:括号、函数、乘方(幂)、乘或除、加或减的顺序,同一优先级的位于左面的先算。

## 2. 比较算符与比较表达式

比较算符建立的是两个量之间的比较关系,由系统判断关系是否成立。如果比较关系成立,比较表达式的值为逻辑值“真”,否则为“假”。相互比较的两个量必须类型一致。无论进行比较的两个量是字符型还是数值型,比较的结果均是逻辑型常量。比较算符表中列出的比较算符均有两种表示方法。表 1-3 括号中的比较算符与括号前的算符是等价的。例如, $A>3$  和  $A \text{ GT } 3$  是等价的。如果  $A=4$  则表达式  $A>3$  为真,其值为 1;如果  $A=3$ ,则表达式  $A>3$  的值为假,其值为 0。

表 1-3 SPSS 基本运算符

数学运算操作符	关系运算符	逻辑运算符
+: 加	<(LT) : 小于	&(and) : 与
-: 减	>(GT) : 大于	(or) : 或
*: 乘	<= (LE) : 小于等于	~ (not) : 非
/: 除	>= (GE) : 大于等于	
** : 幂	= (EQ) : 等于	
( ): 括号	~= (NT) : 不等于	

## 3. 逻辑运算符与逻辑表达式

逻辑运算符即布尔运算符。表 1-3 中,括号前的运算符与括号中的运算符等价,例如  $A \& B$  与  $A \text{ and } B$  是等价的。逻辑运算符与逻辑型的变量或其值构成逻辑表达式。逻辑表达式的值为逻辑型常量。

(1) 与运算:  $\&$  (或 and) 前后的两个量均为真时,逻辑表达式的值为“真”,否则为“假”。

例如,逻辑表达式  $A>B \& C>0$ ,当  $A$  的值大于  $B$  的值且  $C$  为正数时,该逻辑表达式的值为“真”。如果  $A=3$ 、 $B=2$ 、 $C=-6$ ,则该逻辑表达式的值为“假”。

(2) 或运算:  $|$  (或 or) 前后的两个量只要有一个为“真”时,逻辑表达式的值即为“真”。只有当运算符前后两个量均为假时,逻辑表达式的值才为“假”。

例如逻辑表达式  $A>B | C>0$  只要  $A>B$  成立,无论  $C$  为何值,表达式的值均为“真”。或者只要  $C>0$  成立,无论  $A$  与  $B$  为何值,该表达式的值也为“真”。当  $A<B$ ,同时  $C\leq 0$  时该逻辑表达式的值为“假”。

(3) 非运算:  $\sim$  (或 NOT) 是前置运算符,它对其后面的量作非运算。NOT 后面的量值为“真”,则 NOT 运算结果为“假”;后面的量值为“假”,NOT 运算的结果为“真”。

例如,逻辑表达式  $\text{NOT } (A>0)$ ,  $A$  为正数,逻辑表达式的值为“假”;  $A$  为负数或  $A$  为 0,逻辑表达式的值均为“真”。

1.5.3 概率事件

在数据编辑器的 Data View 窗口中是个二维表格。每行都是数据文件的一个记录，在统计学中称作“一个概率事件”，在 SPSS 的菜单中或帮助信息中用“Case”这个单词表示。每个 Case 由各变量的一定的值组成，是对一个事件，或者说是由一个被观测对象的各种特征的实测值或派生值组成的。因此相对“变量”来说可以称之为“观测量”或称为“观测”。单元格中的数值既是某个变量值也是某个观测量中的一个值，因此可以称之为××变量值，也可以称之为某个观测量的某个变量值。

1.5.4 SPSS函数

SPSS 有 18 类函数，见表 1-4。函数的表示方法是在函数关键字后面括号中写入函数自变量。函数自变量有的要求使用单个值或变量名，有的要求使用“:”隔开的多个变量名，还有的允许使用表达式。当然，如果使用变量名或带有变量名的表达式作为自变量，则必须在使用该函数之前对这些变量赋值。下面列出 SPSS 函数。函数类型即函数值的类型。

函数中使用的符号说明：numexpr 数值型表达式；radians 以弧度为单位的角度。

表 1-4 SPSS 函数类型清单

序号	类 型		数量
1	Arithmetic	算术函数	13
2	CDF & Noncentra CDF	累积分布函数	30
3	Conversion	转换函数	3
4	Current Date/Time	当前日期、时间函数	4
5	Date Arithmetic	日期算术函数	3
6	Date Creation	日期生成函数	6
7	Date Extraction	日期提取函数	11
8	Inverse DF	反分布函数	18
9	Miscellaneous	混杂函数	4
10	Missing Values	缺失值函数	6
11	PDF & Noncentra PDF	概率密度函数	27
12	Random Number	随机数函数	22
13	Search	查找函数	10
14	Significance	显著性函数	2
15	Statistical	统计函数	7
16	String	字符函数	26
17	Time Duration Creation	时间间隔生成函数	4
18	Time Duration Extraction	时间间隔提取函数	8

1. 算术函数（Arithmetic）13 个

(1)ABS (numexpr) 数值型函数，函数值为数值表达式的绝对值。

(2) **ARSIN** (*numexpr*) 数值型函数, 函数值为数值表达式的反正弦值, 单位为弧度, 自变量 *numexpr* 其范围在-1~1 之间。

(3) **ARTAN** (*numexpr*) 数值型函数, 函数值为数值型自变量表达式 *numexpr* 的反正切值, 单位为弧度。

(4) **COS** (*radians*) 数值型函数, 函数值为单位为弧度的自变量表达式 *radians* 的余弦值。

(5) **EXP** (*numexpr*) 数值型函数, 函数值为以 *e* 为底, 以括号中的自变量表达式 *numexpr* 值为指数的幂值。应该注意, 若指数太大或函数值太大, 其结果会超出 SPSS 的计算范围。

(6) **LN** (*numexpr*) 数值型函数, 函数值为以 *e* 为底的自然对数值, 自变量数值表达式 *numexpr* 必须是数值型, 而且要大于 0。

(7) **LNGAMMA**(*numexpr*) 数值型函数。函数值为数值表达式 *numexpr* 的完全 Gamma 函数的对数。表达式必须是数值型的, 且其值必须大于 0。

(8) **LG10** (*numexpr*) 数值型函数, 函数值为以 10 为底的对数值, 数值表达式 *numexpr* 必须是数值型, 而且值要大于 0。

(9) **MOD** (*numexpr, modulus*) 数值型函数, 函数值为数值表达式 *numexpr* 除以模数 *modulus* 的余数。两个自变量必须是数值型, 模数不能是 0。

(10) **RND** (*numexpr*) 数值型函数, 函数值为数值表达式 *numexpr* 的值取四舍五入后的整数。

(11) **SIN** (*radians*) 数值型函数, 自变量 *radians* 是以弧度为单位的角度, 函数值为弧度角的正弦值。

(12) **SQRT** (*numexpr*) 数值型函数, 函数值为一个正数的平方根。数值表达式 *numexpr* 的值必须大于等于 0。

(13) **TRUNC** (*numexpr*) 数值型函数, 函数值为数值表达式 *numexpr* 的值被截去小数部分的整数。

2. 累积分布函数 (CDF & Noncentral CDF) 30 个, 详见第 4 章。

3. 转换函数 (Conversion) 3 个

(1) **NUMBER** (*strexpr, format*) 数值型函数, 当字符串内容为一串数字时, 该函数返回字符串表达式作为数字的值, 返回的函数值可以参与运算。第二个表达式为一个数值格式, 用来读取字符串表达式中的数字。

如果 *name* 是一个由 8 个数字组成的字符串, **NUMBER**(*name, f8*) 就是由这些数字表示的数值。如果字符串不能使用指定的格式, 该函数返回系统缺失值。

(2) **NUMBER**(*stringDate, DATE11*) 数值型函数, 把内容为标准格式 (dd-mmm-yyyy) 日期的字符串 转换成描述该日期的秒数。如果字符串不能使用标准格式读取, 函数值是系统缺失值。第一个自变量是字符型, 自变量的值为与 *Date11* 格式相应的日期。



如果我们定义了字符串格式的自变量,输入了与 dd-mmm-yyyy 相应的日期,可以使用该函数将字符串变量转换为日期变量。

(3) **STRING (numexpr,format)** 字符型函数,根据 *format* 所设定的格式将数值表达式转换为字符串。例如 `string(-1.5,F5.2)` 返回字符串 ‘-1.50’。第二个自变量 *format* 必须是一个数值的格式。

**注意: 数值与数字有区别, 以上所讲的数值是数, 数字指的是表现为数字的字符。**

4. 当前日期/时间函数 (Current Date/Time) 4 个。

5. 日期算术函数 (Date Arithmetic) 3 个。

6. 日期生成函数 (Date Creation) 6 个。

7. 日期提取函数 (Date Extraction) 11 个。

有关日期的函数和应用见第 4 章。

8. 反分布函数 (Inverse DF) 18 个, 详见第 4 章。

9. 混杂函数 (Miscellaneous) 4 个。

(1) **\$Casenum** 当前观测量的顺序号。对每个 case, **\$Casenum** 是读取的并包括这个观测的观测量号, 格式是 F8.0。 **\$Casenum** 的值不一定是数据编辑窗中的行号, 如果文件排序或者新的观测代替了文件末尾之前的观测, 这个值也会改变。

(2) **LAG(variable)** 数值型或字符型函数。函数值是前一个观测的变量值。

(3) **LAG(variable[, n])** 数值型或字符型函数。函数值是前一个或前 *n* 个观测的变量值。第 2 个自变量是可选的。*n* 必须是正整数; 默认值为 1。例如 `prev4=LAG(gnp,4)` 的值为当前观测之前的第 4 个观测的变量 **gnp** 的值。

(4) **VALUELABEL(varname)** 字符型函数。函数值是变量值的标签, 当该值没有标签时函数值是空字符串。自变量 **varname** 必须是变量名, 不能是表达式。

10. 缺失值函数 (Missing Values) 6 个

(1) **\$SYSMIS** 数值型函数, 产生系统缺失值。常用于判断并记录缺失值。例如如果取得的数据中有小于 1.4m 的观测, 而身高 < 1.4m 就不能参与一项研究。可以执行语句:

```
IF (height<1.40) height=$Sysmis.  
EXECUTE.
```

就可将身高变量值小于 1.4 的身高值改为圆点。可以在 **transform**→**compute variable** 打开相应对话框完成操作。

(2) **MISSING (variable)** 逻辑型函数, 如果变量具有缺失值, 返回 1 或者 true。自变量应该是工作数据文件中的变量名。

(3) **NMISS (variable [,...])** 数值型函数, 函数值是自变量表中各自变量具有的系统缺失值或读者缺失值的总数。此函数需要至少一个自变量, 这些自变量必须是当前工作数据文件中的变量名。

(4) **NVALID(variable[...])** 数值型函数。函数值为自变量表中的变量具有的合法的非缺失值的总数。函数要求至少一个自变量, 自变量应该是当前工作数据文件中的变量名。

(5) **SYSMIS (numvar)** 逻辑型函数, 如果 *numvar* 的值为系统缺失值, 函数值为 1 或者 **true**。自变量 *numvar* 必须是工作数据文件中的一个数值型变量的变量名。

(6) **VALUE (variable)** 数值型或字符型函数, 忽略读者定义的缺失值, 返回变量值。自变量必须是工作数据文件中的变量名。

需要说明的是, 函数和简单的算术表达式用不同的方法处理缺失值。

① 在表达式  $(var1+var2+var3)/3$  中, 如果一个观测量的三个变量中任意一个是缺失值, 运算结果就是缺失值。

② 在表达式 **MEAN(var1, var2, var3)** 中, 仅当一个观测量的所有变量的值都是缺失值时, 运算结果才是缺失值。

③ 对于统计函数, 可以在函数名后面, 指定非缺失值的最小数。为此, 在函数名后面打一个逗号, 以及至少要有非缺失值数目, 例如 **MEAN.2(var1, var2, var3)**。

11. 概率密度函数 (PDF& Noncentral PDF) 27 个, 详见第 4 章

12. 查找函数 (Serch) 10 个 (与其他类拆分的有 8 个)

(1) **ANY (test, value [, value...])** 逻辑型函数, 如果 *test* 的值与其后的 *value [, value,...]* 中的某一数值匹配, 那么函数值为真, 返回 1 或 **True**; 否则, 函数值为假, 返回 0 或者 **False**。这个函数需要至少两个自变量。该函数要求至少两个自变量。例如 **ANY(var1, 1, 3, 5)**, 如果 *var1* 的值是 1 或 3 或 5, 函数值为 1, *var1* 为其他值, 函数值为 0。该函数还可以用来在变量表或表达式表中扫描一个值。例如 **ANY(1, var1, var2, var3)** 如果在三个指定的变量中任意一个变量有 1 值, 函数值为 1; 所有变量的值都不是 1, 函数值为 0。

(2) **RANGE (test, lo, hi [, lo, hi,...])** 逻辑型函数, 如果 *test* 的值包含在由 *lo, hi* 所定义的范围内, 函数值为 1 或者 **true**, 否则为 0 或者 **False**。所有变量必须都为数值型或都为字符型, 并且所设置的 *lo, hi* 变量的大小顺序必须为  $lo \leq hi$ 。注意, 不同地区使用不同语言, 对自变量为字符型的情况, 同一个函数运算结果可能有很大区别。本函数按 ASCII 码顺序运算。

另有 6 个字符串函数: **CHAR.INDEX(2)**、**CHAR.INDEX(3)**、**CHAR.RINDEX(2)**、**CHAR.RINDEX(3)**、**REPLACE(3)**、**REPLACE(4)**。重复出现在字符串函数类中。

另外 SPSS 16.0 把 **Max**、**Min**、**Range** 也列入了查找函数。前面两个在统计函数中重复出现。在这里不再解释。

13. 显著性函数 (Significance) 2 个

14. 统计函数 (Statistical) 7 个

(1) **CFVAR (numexpr,numexpr[...])** 数值型函数。函数值为自变量 (或数值表达式 *numexpr* 的值) 的变异系数 (标准差除以均值)。此函数要求两个或两个以上的自变量。自变量必须为数值型, 而且必须有合法值。

(2) MAX (*value,value[,...]*) 数值型函数或字符型函数, 函数值为自变量 *value* 所有合法值的最大值。至少需要两个以上 *value*。

(3) MEAN (*numexpr,numexpr[,...]*) 数值型函数, 函数值为多个数值表达式 *numexpr* 的算术平均数。数值表达式至少需要两个以上。

(4) MIN (*value,value[,...]*) 数值型函数或字符型函数, 函数值为具有合法值的自变量 *value* 的最小值。至少需要两个以上的 *value*。

(5) SD (*numexpr,numexpr[,...]*) 数值型函数, 函数值为所有数值表达式标准差。这个函数需要两个或两个以上的自变量, 自变量可以是表达式, 或者为非缺失的合法值, 而且必须为数值型。

(6) SUM (*numexpr,numexpr[,...]*) 数值型函数, 函数值为所有数值表达式值的累加和。这个函数需要两个或两个以上的非缺失合法值。自变量可以是数值、数值型表达式。

(7) VARIANCE (*numexpr,numexpr[,...]*) 数值型函数, 函数值为所有数值表达式的方差。这个函数需要两个或两个以上的自变量。自变量可以是表达式, 但必须是数值型。

## 15. 字符串函数 (String) 26 个

(1) CHAR.INDEX (*haystack,needle*) 数值型函数, 返回一个整数, 它表明 *needle* 代表的字符串在 *haystack* 代表的字符串中第一次出现的起始位置。如果返回值为 0, 表明字符串 *needle* 不存在在字符串 *haystack* 中。在函数表中, 该函数名称为 CHAR.INDEX(2), 意为两个自变量。CHAR.INDEX(*var1*, 'abcd')将返回整个字符串“abcd”在字符串变量 *var1* 的起始位置。函数列表中该函数的函数名为 CHAR.INDEX(2)。

(2) CHAR.INDEX (*haystack,needle,divisor*) 数值型函数, 见前一个函数。其第三个自变量 *divisor* 是可选择的, 它必须是一个整数, 表明将字符串 *needle* 均匀地分为要被查询的独立的子字符串的字符数。例如 CHAR.INDEX(*var1*, 'abcd', 1)返回字符串中任意一个字符在字符串变量 *var1* 代表的字符串中第一次发生的位置。CHAR.INDEX(*var1*, 'abcd', 2)返回的值是“ab”或“cd”在字符串中第一次发生的位置。*divisor* 必须是正整数, 必须把 *needle* 分成均匀的长度。*needle* 或子串在 *haystack* 中不存在, 函数值为 0。函数列表中该函数的函数名为 CHAR.INDEX(3)。

(3) CHAR.LENGTH(*strexpr*)数值型函数, 函数值为自变量 *strexpr* 值的以字符为单位, 去掉尾部空格后的长度。

(4) CHAR.LPAD (*strexpr,length*) 字符型函数, 返回一个字符串, 在字符串表达式的左侧增加空格扩展到 *length* 所规定的长度。*length* 必须是正整数, 其范围从 1~255。在函数列表中, 此函数在函数列表中名为 CHAR.LPAD(2)。

(5) CHAR.LPAD (*strexpr,length,char*) 字符型函数。与前一个相同, 但是, 不是用空格而是用 *char* 变量代表的字符串的完整复制在 *strexpr* 代表的字符串左侧扩展。*char* 必须是用单引号括起的字符串常量。此函数在函数列表中名为 CHAR.LPAD(3)。

(6) CHAR.MBLEN(*strexpr,pos*) 数值型函数, 返回字符表达式 *strexpr* 代表的字符在

*pos* 指定的位置上的字符所占的字节数。

(7) CHAR.RINDEX (*haystack,needle*) 数值型函数, 返回一个整数, 它表明字符串 *needle* 在字符串 *haystack* 中最后出现的开始位置。返回 0 表示字符串 *needle* 不在 *haystack* 中。例如 CHAR.RINDEX (*var1*, 'abcd') 返回整个字符串 “abcd” 在自变量 *var1* 的值代表的字符串中最后一次出现的位置。此函数在函数列表中名为 CHAR. RINDEX (2)。

(8) CHAR.RINDEX (*haystack,needle,divisor*) 数值型函数, 返回一个整数, 它表明字符串 *needle* 在字符串 *haystack* 中最后出现的开始位置。返回 0 表示字符串 *needle* 不在 *haystack* 中。第三个自变量是可选择的, 它是一个整数, 用来表示将字符串 *needle* 平均分成被查询的字符串的数目。它必须是一个可以将字符串 *needle* 整分的正整数。没有第 3 个自变量, 功能与上一个函数相同。此函数在函数列表中名为 CHAR. RINDEX (3)。

例如, CHAR.RINDEX(*var1*, 'abcd', 1) 最后参数 1 表明, 把 “abcd” 分成单独的一个个的字符, 函数值为任何一个字符在自变量 *var1* 值代表的字符串中最后一次出现的位置。

CHAR.RINDEX(*var1*, 'abcd', 2) 最后参数 2 表明, 把 “abcd” 分成长度相等的两部分 “ab” 和 “cd”, 函数值是这两个字符串中任何一个在自变量 *var1* 值代表的字符串中最后一次出现的位置。此函数在函数列表中名为 CHAR. RINDEX (3)。

(9) CHAR.RPAD (*strexpr,length*) 字符型函数, 返回字符串, 它的长度由 *length* 决定: 在字符串表达式的右侧加空格, 以达到 *length* 的长度, *length* 的值必须在 1~255 之间。此函数在函数列表中名为 CHAR. RPAD(2), 是两个自变量的函数。

(10) CHAR.RPAD (*strexpr,length,char*) 三个自变量的字符型函数, 返回字符串。第三个变量 *char* 是可选的, 没有第 3 个自变量, 函数功能与上一个函数相同。函数值是在字符串的右侧增加若干自变量 *char* 代表的字符, 达到自变量 *length* 指定的长度。*char* 必须是一个带有引号的单个字符或其值是单个字符的字符表达式。此函数在函数列表中名为 CHAR. RPAD(3)。

(11) CHAR.SUBSTR (*strexpr,pos*) 字符型函数, 函数值为自变量 *strexpr* 代表的字符串中从 *pos* 开始到其结尾处的子字符串。此函数在函数列表中名为 CHAR.SUBSTR(2)。

(12) CHAR.SUBSTR (*strexpr,pos,length*) 字符型函数, 函数值为自变量 *strexpr* 代表的字符串中从 *pos* 开始, 长度为 *length* 的子字符串。此函数在函数列表中名为 CHAR.SUBSTR(3)。

(13) CONCAT (*strexpr, strexpr [...]*) 字符型函数, 函数中每个自变量都是一个字符串表达式。该函数返回一个字符串, 它是各自变量代表的字符串按括号中的顺序串接起来的结果。此函数要求两个或两个以上的字符型自变量。

(14) LENGTH(*strexpr*) 数值型函数, 返回 *strexpr* 代表的字符串长度。对于 Unicode 码的字符串变量, 它是每个自变量值, 包括尾部空格的字符数, 但对编码页面模式, 它就是定义的, 包括尾部空格的变量长度。在编码的页面模式下, 要得到以字节为单位的

除去尾部空格的长度，需要使用嵌套函数调用 `LENGTH(RTRIM(strexp))` 来求得。

(15) `LOWER(strexp)` 字符型函数，函数值为将自变量 `strexp` 中的大写字母改变为小写字母，其他字符不变。自变量可以是字符串变量、字符串表达式，也可以是字符串常量。例如变量 `name` 的值是 `Jery`，`LOWER(strexp)` 的值为 `jery`。

(16) `LTRIM(strexp)` 字符型函数，函数值为自变量 `strexp` 值去掉首部空格的结果。在函数列表中的函数名为 `LTRIM(1)`。

(17) `LTRIM(strexp[,char])` 字符型函数，函数值为自变量 `strexp` 的值去掉首部变量 `char` 值代表的字符。第 2 个自变量 `char` 的值必须是单个字符。在函数列表中的函数名为 `LTRIM(2)`。

(18) `MBLEN.BYTE(strexp,pos)` 数值型函数。函数值是自变量 `strexp` 在 `pos` 指定位置以字符为单位的字节数（如英文字符是 1 个字节，中文是 2 个字节）。

(19) `NORMALIZE(strexp)` 字符型函数，函数值是自变量 `strexp` 的规范化版本在 Unicode 模式中函数值是 Unicode NFC。对页面方式无效，函数值就是自变量值，但长度可能与输入的长度不同（Unicode 国际统一编码标准）。

(20) `NTRIM(varname)` 函数值是自变量 `varname` 的没有去掉尾部空格的值，自变量 `varname` 的值必须是一个变量名，不能是个表达式。

(21) `REPLACE(a1,a2,a3)` 字符型函数。在 `a1` 代表的字符串中的所有 `a2` 字符串都用 `a3` 字符串代替。自变量 `a1`、`a2`、`a3` 必须在函数调用前处理成字符串值。例如 `REPLACE("abcabc", "a", "x")` 函数值为 `"xbcxbc"`。在函数列表中该函数名为 `REPLACE(3)`。

(22) `REPLACE(a1,a2,a3[,a4])` 字符型函数。在 `a1` 代表的字符串中的 `a2` 字符串用 `a3` 字符串代替 `a4` 次。可选的自变量 `a4` 指定替换发生的次数。自变量 `a1`、`a2`、`a3` 必须在函数调用前处理成字符串值（字符串变量或者括在引号中的字符串常量）。可选的自变量 `a4` 必须处理成非负整数。例如 `REPLACE("abcabc", "a", "x", 1)` 函数值为 `"xbcab"`。在函数列表中该函数名为 `REPLACE(4)`。

(23) `RTRIM(strexp)` 字符型函数，返回截取了尾部空格后的字符串。该函数在函数列表中名为 `RTRIM(1)`。

(24) `RTRIM(strexp,char)` 字符型函数，函数值是自变量 `strexp` 的值截取了尾部 `char` 代表的字符后的字符串。`char` 必须是一个带有引号的单个字符或其值是单个字符的字符表达式。该函数在函数列表中名为 `RTRIM(2)`。

(25) `STRUNC(strexp,length)` 字符型函数。函数值是自变量 `strexp` 截取 `length` 指定的长度（字节为单位）然后去掉尾部空格。

(26) `UPCAS(strexp)` 字符型函数，函数值为字符串表达式 `strexp` 值中小写字母变为大写后的字符串。

16. 时间间隔生成函数（Time Duration Creation）4 个，见第 5 章。

17. 时间间隔提取函数 (Time Duration Extraction) 8 个, 见第 5 章。

## 1.6 获得帮助

### 1.6.1 SPSS帮助系统

单击各窗口的 **Help** 就可以展开系统帮助菜单, 如图 1-29 所示, 可获得多项帮助。不同窗口的帮助菜单内容略有不同, 主要的帮助内容如下。

(1) 单击 **Help**→**Topics** 打开 **Online Help** 对话框, 如图 1-30 所示。在目录选项卡、索引选项卡和搜索选项卡中, 可以分别按目录查找、输入关键字按索引查找, 或输入单词进行搜索, 类似 Windows 系列软件的帮助系统, 不再赘述。

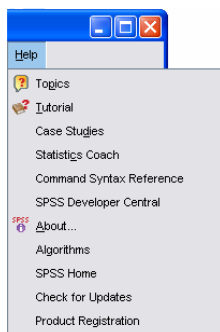


图 1-29 系统帮助菜单

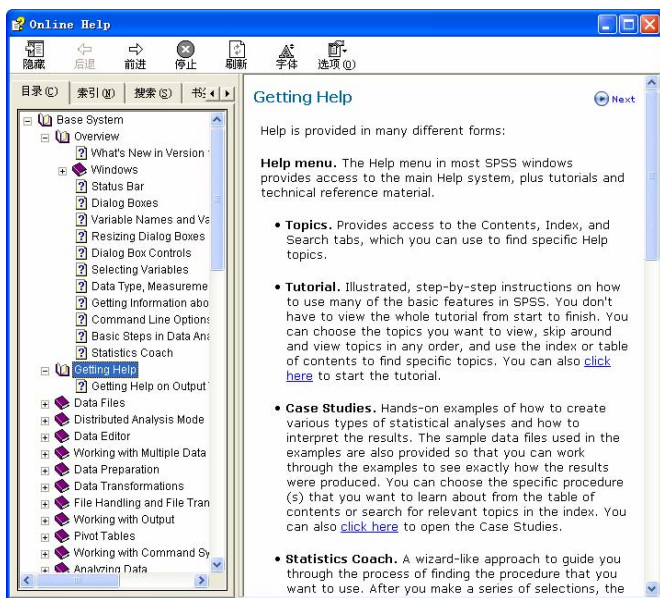


图 1-30 Online Help 对话框

(2) 单击 **Help**→**Tutorial** 打开 **Tutorial** 帮助系统, 对初学者是入门的向导。见图 1-31。在图 1-31(a)中, 单击十字图标, 可以一层层打开树形目录。单击一个菜单项, 可以获得如图 1-31(b)的指导画面。窗口分两半。左面的是用图解释, 右面是文字说明。单击右下角的左右箭头按钮可以向后翻页、向左按钮向前翻页; 单击小房子图标按钮回到左图的目录, 可以按目录查找需要的帮助信息; 单击放大镜按钮, 可以输入关键字, 搜索需要的帮助信息。

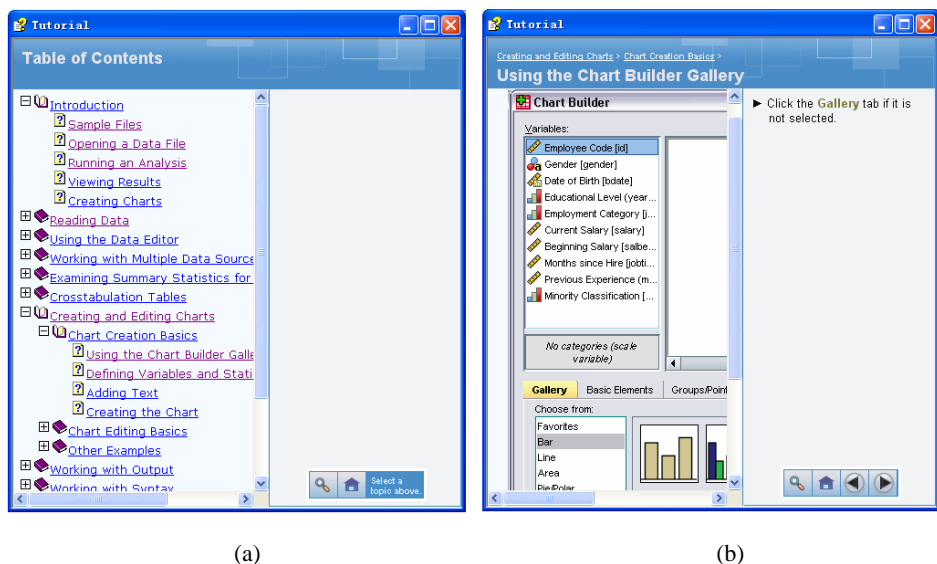


图 1-31 Tutorial 的菜单和指导内容示例

(3) 单击 **Help**→**Case Studies** 进入另一个 Tutorial 对话框, 是对各种分析过程的操作指导, 内容很丰富, 有例题, 有操作步骤, 有选项说明, 还有输出结果的解释及结论的得出。操作方法与 Tutorial 相同。这个菜单只有在数据窗口和输出窗口的 **Help** 菜单中才有。

(4) 单击 **Help**→**Statistics Coach** 打开统计学指导系统, 是对基本统计方法的指导。

指导内容按照下面的思路和结构解决读者在初步学习统计分析时可能出现的问题。即: 要分析什么、数据是什么类型、要什么样的输出、操作步骤以及对话框的详细说明。

在窗口中, 左面窗口保持树形目录, 便于查找。右面窗口根据需要指导的内容分两列列出读者可能存在的问题。

① **What do you want to do?** 左列列出要进行的分析内容, 供读者选择, 右面一列对应左面的项目为 **Show example output**, 一级窗口要求选择要进行的分析, 列出了指导的主要内容, 见图 1-32(a)。单击目录窗口中树形结构中的一项目, 或者在右窗口左列中选择一个主题, 打开如图 1-32(b)所示的帮助窗口。可以单击一项标题, 然后单击 **Next** 按钮看其解题过程, 单击 **Show example output**, 可以看到例题输出。对选择的每一项分析, 都可以单击 **Next** 按钮进入下面各级对话框。

② **What kind of data do you want to (或 What kind of data do you have)** ……典型的二级窗口, 列出可以分析的数据类型, **Help** 对各种类型数据进行进一步的阐述; 在左列显示各种可能的数据类型。选择一种数据类型, 单击 **Next** 按钮, 显示对数据的说明; 单击 **Show example output**, 针对指定的数据类型, 给出例题输出。

③ **What kind of display do you want?** 典型的三级窗口, 列出可能的输出, 是表格还

是图形? Help 对各项输出做详细的说明, 单击 Show example output, 给出例题输出。

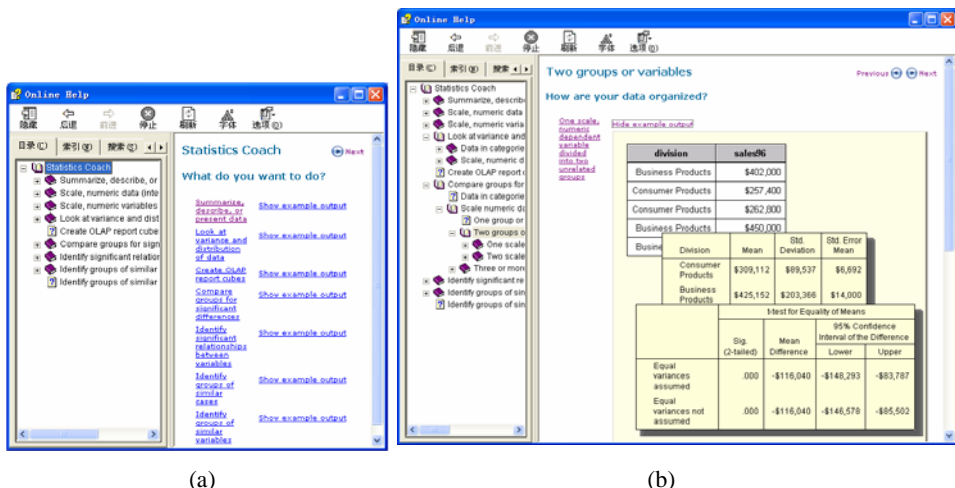


图 1-32 Statistics Coach 的一级窗口及使用统计学指导的帮助窗口

(5) 关于过程语句的帮助系统。SPSS 窗口操作方式使操作变得容易, 但是包含的方法和选项有限。需要使用语句补充分析功能和窗口运行方式没有包括的分析过程。对高级分析方法, 语句的帮助信息就显得更重要。语句帮助信息显示在 Adobe Reader 阅读器窗口中, 见图 1-33。

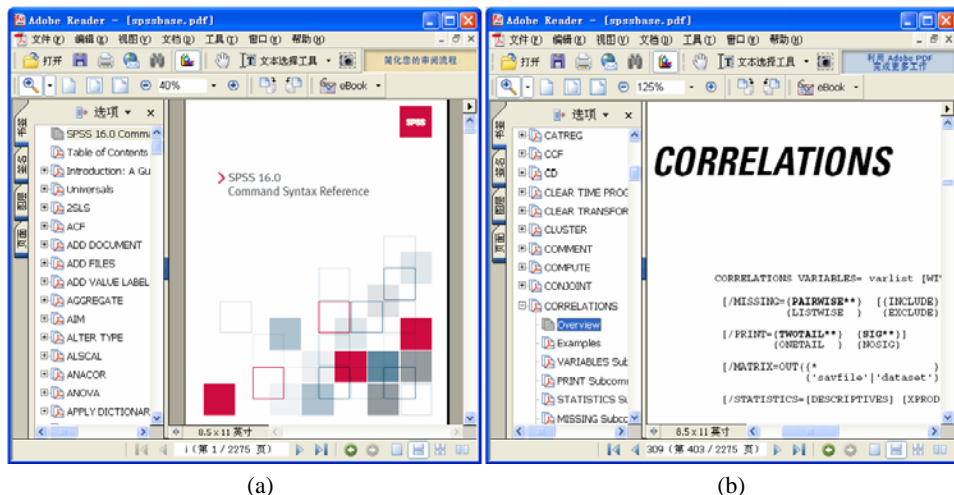


图 1-33 语句帮助窗口

说明:

(1) 因为语句的帮助文件是 PDF 格式的文件, 阅读语句的帮助信息需要安装 Acrobat



Reader。

(2) 语句的帮助文件很大，对于硬盘容量较小的计算机，不必安装语句帮助文件。可以直接从光盘读取。

操作方法是单击 **Help→Command Syntax Reference** 自动打开 **Acrobat Reader**，语句帮助信息显示在 **Acrobat Reader** 阅读器窗口中，见图 1-33 (a)。

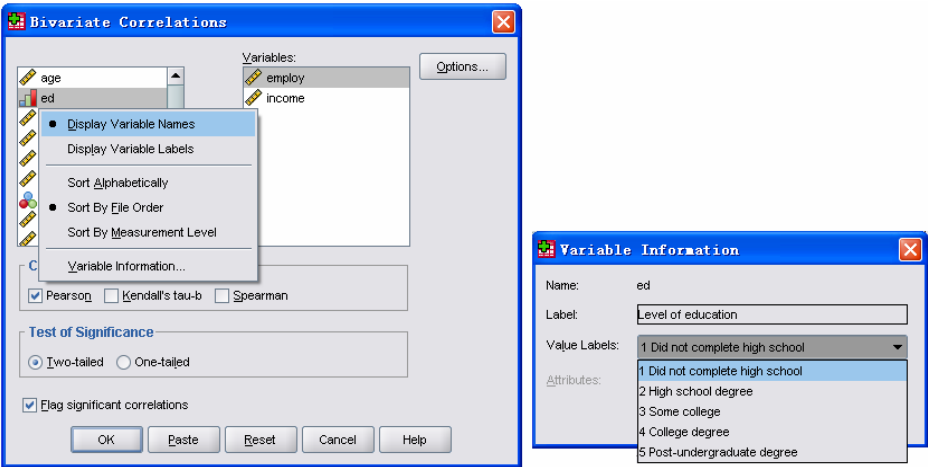
窗口左边是语句帮助信息的菜单，单击一个具体的过程语句名，右窗口显示具体的语句帮助信息，见图 1-33 (b)。

1. 6. 2 右键帮助

1. 对话框中的右键帮助

在对话框的变量表中，用右键单击一个变量，出现小菜单，见图 1-34 (a)，包括以下 3 组 6 项。在小菜单中组间用横线隔开。

- (1) **Display Variable Names** 显示变量名。
- (2) **Display Variable Labels** 显示变量标签。变量表显示变量名还是变量标签，在 **Options** 对话框中设置。在这里可以改变。
- (3) **Sort Alphabetically** 按字母顺序排列变量。
- (4) **Sort By File Order** 按变量在数据文件中出现的顺序排列。
- (5) **Sort By Measurement Level** 按测度水平排列。改变变量的排列顺序，便于查找。
- (6) **Variable Information**，选择这一项，打开 **Variable Information** 对话框，给出变量详细信息，包括值标签的下拉列表，见图 1-34(b)。这些帮助信息有助于选择分析变量。



(a)

(b)

图 1-34 对话框中变量的右键帮助

## 2. 输出表格中的右键帮助

在输出窗口中双击一个表格，激活它。在表格输出的某个统计量上，单击右键都会出现一个列表，见图 1-35(a)，选择第一项“**What's This?**”，会给出对该项统计量的解释，见图 1-35(b)。

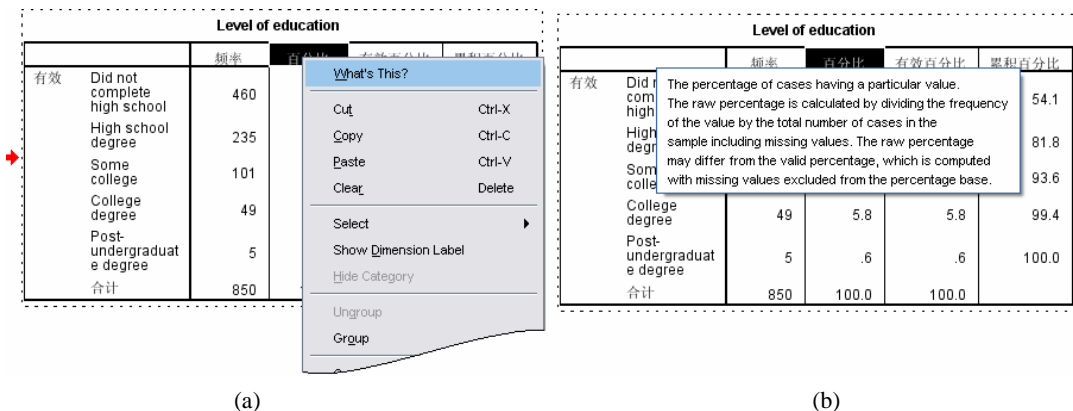


图 1-35 输出表格的右键帮助

## 习 题 1

1. SPSS 软件有几种运行方式？什么是混合运行方式，它有什么特点？
2. SPSS 有几种类型的窗口，每个窗口的主要功能是什么？
3. 什么是输出窗口（或语句窗口）的主窗口，什么是主窗口的标志？怎样把非主窗口变成主窗口？分别叙述主窗口和非主窗口的作用，以输出窗口为例说明之。
4. 通过什么菜单项设置系统参数？
5. SPSS 的统计分析功能分布在何处？
6. 从何处可以获得帮助信息？系统提供的帮助有几种形式？

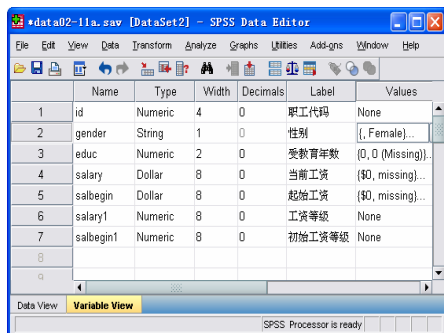
## 第2章 数据与数据文件

### 2.1 变量定义与数据编辑

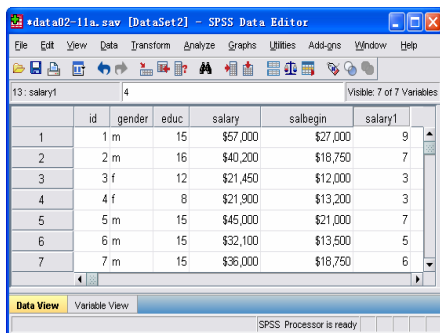
#### 2.1.1 数据编辑器

SPSS 启动后，屏幕上出现“Data Editor”窗口，这是 SPSS 的数据编辑器，也称数据窗口，如图 2-1 所示。读者在该窗口可建立、打开数据文件。为便于建立变量和查看变量属性，数据窗口分为：Data View、Variable View 两种，其组成与功能如下：

1. 窗口标题栏：当 SPSS 启动后，屏幕显示窗口名称为“Untitled-SPSS Data Editor”。当窗口中数据已经保存到数据文件时，窗口标题栏则显示该窗口中数据文件名。窗口标题栏下面是菜单栏和工具栏。



(a)



(b)

图 2-1 数据编辑器的两个窗口

2. Variable View 窗口用于定义和编辑变量的属性，见图 2-1(a)。变量显示区是一个二维表格。左面是行号，也是变量的序号。变量的属性显示在平面表格的第一行，包括变量的变量名、类型、宽度、小数位数、变量标签、值标签、显示格式和测度方式等。

3. Data View 窗口用于显示和编辑数据，见图 2-1(b)。

在工具栏下面是数据栏与数据输入栏：左边一栏是当前数据栏，显示当前光标位置上的变量名和当前记录号。右边一栏为数据输入栏，显示光标位置上的数据值。从键盘输入的数据先显示在此栏中，回车后系统根据定义的变量长度选择合适的形式显示在光标定位的单元格中。数值过大或过小，都有可能使用科学记数法显示数据。

数据显示区：是一个二维平面表格，左面的行号即观测量序号；在表格顶部显示变量名，在它下面的各单元格中显示各变量值。被选定的单元格边框色加深，单元格有底色。所选定单元格中的数据值显示在数据输入栏中。

### 2.1.2 定义变量

输入数据之前首先要定义变量。定义变量即要定义变量名、变量类型、变量长度（小数位数）、变量标签（或值标签）和变量的格式（显示宽度、对齐方式、缺失值标记等）、缺失值和测度方式。

定义变量的步骤如下：

1. 单击 **Variable View** 选项卡，使数据编辑窗口置于定义变量状态，如图 2-2 所示，每行定义一个变量。

2. 定义变量名

光标置于 **Name** 列的空单元格中，单击单元格后输入变量名。例如输入 **Gender** 作为变量名。回车后在同行各单元格中系统自动给出了变量的默认属性。

3. 变量的默认属性值

- **Type**: 变量类型，默认类型为数值型（**Numeric**）。
- **Width**: 变量长度，默认长度为 8。
- **Decimals**: 小数位数，默认小数位数为 2。
- **Label**: 变量标签；**Values**: 值标签；**Missing**: 缺失值，读者自定。
- **Columns**: 列宽，变量在 **Data View** 中所占列宽默认为 8 个英文字符。
- **Align**: 对齐方式，默认右对齐（**Right**）。
- **Measure**: 测度方式，默认为等间隔测度（**Scale**）方式。

如果认为默认的属性与要定义的变量属性不符，可以在同行各属性单元格中设置读者所需要的变量属性。

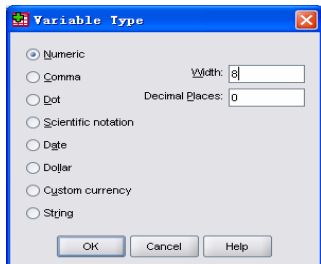


图 2-3 定义变量类型的对话框

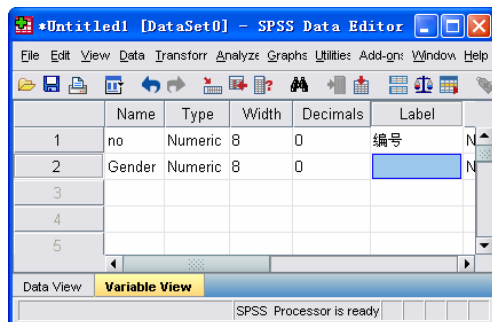


图 2-2 定义变量的窗口

4. 定义变量类型与宽度

(1) 定义变量类型

单击 **Type** 列的单元格，默认值 **Numeric** 旁出现删节号。单击删节号，展开 **Variable Type**（定义变量类型）对话框，如图 2-3 所示。

定义变量类型对话框左半部列有 8 种可供选择的变量类型，自上而下标明的变量类型为：**Numeric**（标准数值型）；**Comma**（带逗点的数值型）；**Dot**（逗点作小数点的数值型）；**Scientific notation**

(科学记数法); Date (日期型); Dollar (带有美元符号的数值型); Custom currency (自定义型); String (字符型)。

单击选择的类型即可。

## (2) 定义变量宽度和小数位数

Width 栏中的数值是变量的总宽度, Decimal 框中显示的是小数位数。要改变其值, 可在单元格中双击鼠标左键, 在编辑状态下输入读者认为合适的值。或者用鼠标单击单元格中出现的上下箭头按钮, 增加或减少变量宽度值。

## 5. 定义变量标签

定义变量标签是为了注释变量名含义。

在 Variable View 窗口中, 双击 Labels 相应的单元格, 输入注释即可, 要尽量简单明了。例如, 对 Gender 变量, 可以给出汉字“性别”作为变量的标签。SPSS16 版和 17 版都可以输入中文标签, 每个分析过程的主对话框的原变量表中会在显示英文变量名的同时显示中文标签, 使操作变得容易。可以使用 Edit 菜单中的 Options 设置在输出表格中是否使用在此定义的中午标签。详见 1.3.6 节中的内容。

## 6. 定义与修改值标签

(1) 定义值标签。单击 Value 栏相应的单元格, 再单击单元格右侧出现的删节号, 打开 Value Labels 对话框, 见图 2-4。在 Value 框中输入变量值, 在 Label 框中输入对该值含义解释的标签。单击 Add 按钮, 一个值标签就被加入到第三个框, 即值标签清单中。例如在定义 Gender 变量过程中, 数值 1 表示男性, 数值 2 表示女性, 则先在 Value 框中

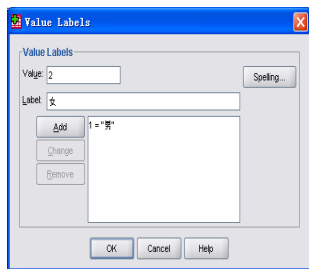


图 2-4 定义值标签的对话框

输入“1”, 在 Label 框中输入“男”, 单击 Add 按钮, 列表框中增加了一个值标签, 显示 1=“男”。用同样方法定义第二个值标签, 清单中显示 2=“女”, 值标签定义完毕。单击 OK 按钮, 确认定义的变量标签和值标签正确无误, 并返回 Variable View 窗口。定义中文值标签并在 Edit 菜单的 Options 功能(见 1.3.6 中的解释)中定义在输出表格中使用这个值标签, 会使解释输出结果变得容易。

(2) 修改值标签。要修改变量的值标签, 在 Value Labels 对话框中, 按如下步骤进行。

首先在值标签列表中选择要加以修改或删除的值标签表达式, 鼠标单击使其反向显示。此时, 变量值和该值的标签分别显示在列表上方的 Value、Label 框中。

删除操作: 单击加亮的 Remove 按钮, 被选定的值标签就从值标签列表中删除。

修改操作: 在 Value 框中可以输入新的变量值, 在 Label 框中输入新标签。例如选择列表中的 Gender 变量值 2 的值标签表达式, 并在 Value 框中修改 Gender 变量值, 将 2 改为 0, 标签“女”不变, 单击 Change 按钮, 在列表中的表达式由 2=“女”改为 0=

“女”，修改完成。

一个值不能定义两个不同的标签；不同的值不能赋予相同的标签。如果用英文定义标签，还可以单击对话框右上角的 **Spelling** 按钮，查拼写错误。

### 7. 定义读者缺失值

在 **Variable View** 窗口中，单击变量与 **Missing** 对应的单元格，然后单击右侧的删节号按钮，展开 **Missing Values** 定义变量读者缺失值对话框，见图 2-5。

先选择一种缺失值的类型，再进行具体定义。定义读者缺失值的类型有 3 种：

(1) **No missing values**，无缺失值。本选项是系统的默认状态。如果当前变量的值测试、记录完全正确，没有遗漏，则可选择此项。

(2) **Discrete missing values**，离散缺失值。选择这种方式，可以在下面的 3 个矩形框中输入 3 个可能出现在相应的变量中的缺失值，也可以少于 3 个。在进行统计分析时系统遇到这几个值，则作为缺失值处理。例如对于性别变量，如果定义了用 1 表示男，用 2 表示女，则值为 0、3、4 都被认为是非法的。可以将这 3 个值分别输入到 3 个矩形框中，当数据文件中出现这几个数据时，系统将按缺失值处理。

(3) **Range plus one optional discrete missing value**，附加一个范围外缺失值。选择此项后，除了 **Low** 和 **High** 参数框外，还有 **Discrete value** 离散值，即范围以外的一个值。例如，如果定义变量 **HEIGHT** 的值中输入的错误数据有 1.40、1.90、1.95 和 2.03，而且在 1.90~2.03 之间没有正确的身高测试值，正确值在大于 1.40 和小于 1.90 的范围内，则可选择此种定义缺失值的方式。在 **Low** 参数框中输入 1.90，在 **High** 参数框中输入 2.03，在 **Discrete value** 参数框中输入 1.40。

如果这三种定义缺失值方式都不能把所有的非法值包括在内，则要在数据文件中查出错误数据进行修改，修改成系统缺失值。或者在 **Syntax** 窗口中利用缺失值函数解决定义缺失值的问题。

### 8. 定义变量的显示格式

#### (1) 定义显示用的列宽度

在 **Variable View** 窗口中，单击 **Columns** 列相应的单元格，再单击出现的上下箭头按钮。增加或减少列宽度值，如图 2-6(a)所示。

#### (2) 定义显示时的对齐方式

在 **Variable View** 窗口中，与变量行 **Align** 列相应的单元格中显示的是默认的对齐方式。对数值型变量，系统默认 **Right**，右对齐；对字符型变量，系统默认 **Left**，左对齐。如果要改变默认的对齐方式，单击 **Align** 列相应的单元格，有 3 种个可选择的方式：**Left**

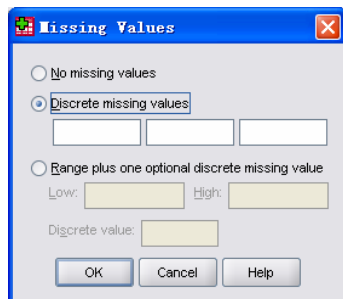


图 2-5 定义读者缺失值对话框

左对齐、Center 居中对齐、Right 右对齐, 在下拉列表中任选一种, 见图 2-6(b)。

### 9. 定义变量测度类型

在 Variable View 窗口中, 与变量行 Measure 列相应的单元格中显示的是默认的变量测度方式 Scale。

#### (1) 关于默认值

- ① 字符串 (字母数字) 变量默认类型为 Nominal, 即标称变量。
- ② 带有值标签的数值型变量默认类型为 Ordinal, 即有序变量或称定序变量。
- ③ 没有定义值标签的数值型变量, 但数值的个数少于指定的数量被设置成 Ordinal。
- ④ 没有定义值标签的数值型变量, 但数值的个数多于 Option1 指定的数量被设置成 Scale, 即等间隔测度的变量。

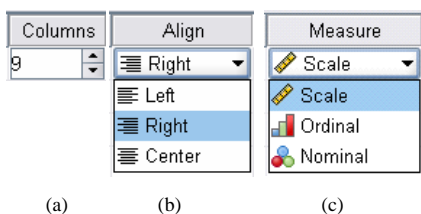


图 2-6 定义变量的列格式和测度方式

默认的变量值的数量为 24。若要改变这个值, 选择 Edit→Options 菜单项, 在对话框中的 Data 选项卡的 Reading External data 栏中设置这个值。

如果要改变默认的测度类型, 单击 Measure 列相应的单元格, 展开下拉列表如图 2-6(c)所示。在下拉列表中有 3 个可选择的类型。

(1) Scale, 尺度变量。对等间隔测度的变量或者表示比值的变量选择此项, 如身高、体重。

(2) Ordinal, 定序变量。对其值表示顺序的变量选择此项, 如比赛名次、职务、职称等, 可以是数值型变量, 也可以是字符型变量。

(3) Nominal, 标称变量。它是分类变量的一种, 可以是数值型变量, 也可以是字符型变量。例如变量值是对所喜欢的颜色的回答, 表示宗教信仰、党派等的变量。

### 10. 确认全部定义的属性

经过上述操作, 定义完一个变量的属性参数。可以重复上述操作, 定义其他变量属性参数。所有变量名及其属性都显示在 Variable View 窗口中。如果对定义的属性满意, 则按 Data View 选项卡, 转移到数据编辑窗口, 输入数据。

## 2.1.3 定义日期变量

定义日期功能可产生周期性的时间序列日期变量, 还可以给时间序列分析的输出加标签。按 Data→Define Dates 顺序展开 Define Dates 定义日期对话框, 如图 2-7 所示。在对话框中选择各项与建立、修改、删除日期型变量有关的操作。

#### 1. 关于 Cases Are 栏

Cases Are 栏即日期类型选项栏, 在其中各项都是定义日期变量的时间间隔和为定义时间变量做准备的功能项。利用该对话框建立具有一定时间间隔的日期变量必须满足下



列条件:

- (1) 在数据窗口中已经有一个数据文件。
- (2) 在该数据文件中的变量名不能与将要建立的日期变量的默认变量名重名, 否则新建日期变量将覆盖同名变量。系统默认变量名有:

YEAR\_、QUARTER\_、MONTH\_、WEEK\_、DAY\_、HOUR\_、MINUTE\_、SECOND\_和DATE\_。

- (3) 对于每个 Cases Are 的功能项, SPSS 生成若干数值型变量, 新变量名以下画线结尾。同时生成一个字符型变量 Date\_用以解释生成的日期变量。例如, 如果选择了 Weeks、days、hours, 则生成四个新变量 WEEK\_、DAY\_、HOUR\_和 DATE\_。

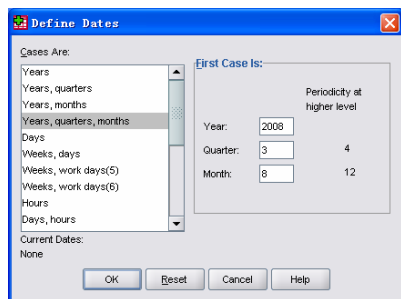


图 2-7 定义日期对话框

## 2. Current Dates 栏

显示项与定义新日期变量有关。在 Current Dates 标题下面显示的是已经存在的与即将生成的日期变量同名的变量及其定义, 以提请注意。

## 3. First Case Is 栏

定义起始日期值, 该值作为第一个观测量, 接下来的各观测量值根据时间间隔自动生成。

## 4. Periodicity at higher level 栏

显示项与 Cases Are 栏目所选择的项目对应, 在 Periodicity at higher level 显示项中指定相应的重复周期。例如一年中的月数, 一周中的天数。在可以输入数值的区域后面显示的是可以输入的最大值。

## 5. Cases Are 栏目中的主要功能

(1) Not dated (倒数第 2 项)。选择此项将删除当前数据文件中与系统默认的日期型变量名相同的变量, 为使用 Cases Are 中的某些功能项的执行创造条件。

(2) Custom (倒数第 1 项)。该功能指出由命令语句生成的日期变量而非使用 Define dates 功能生成的日期变量。例如每周 4 个工作日的日期变量, 它只反映当前工作的数据文件状态, 对数据文件没有影响。

除以上两项功能外全部都是生成日期变量的功能项。

### 【例 1】生成日期变量。

以定义年、季度、月为例说明操作方法。

(1) 按 Data→Define Dates 顺序单击菜单项, 打开 Define Dates 对话框。

(2) 在 Cases Are 栏内选择 Years、quarters、months 项。在 First Case Is 栏内显示:

① Year 框显示 1900, 这是系统默认数值, 改变该值输入 2008, 见图 2-7, 此项表明第一观测量的 Year\_变量值为 2008;

② Quarter 框显示变量的起始值为 1, 周期为 4, 按 1、2、3、4 顺序排列; 输入 3。



③ Month 框显示变量的起始值为 1，周期为 12。输入 8。

单击 OK 按钮，在数据窗口中生成的新变量有：Year\_、Quarter\_、Month\_和对这三个变量值的解释变量 Date\_，如图 2-8 所示。

从图中可以看出，要想使用 Define Dates 功能自动生成日期变量，在原数据文件中的各观测量必须都是按某时间顺序取得的，而且时间顺序必须与 Define Dates 对话框中 Cases Are 栏目中的某一选项相对应才行。

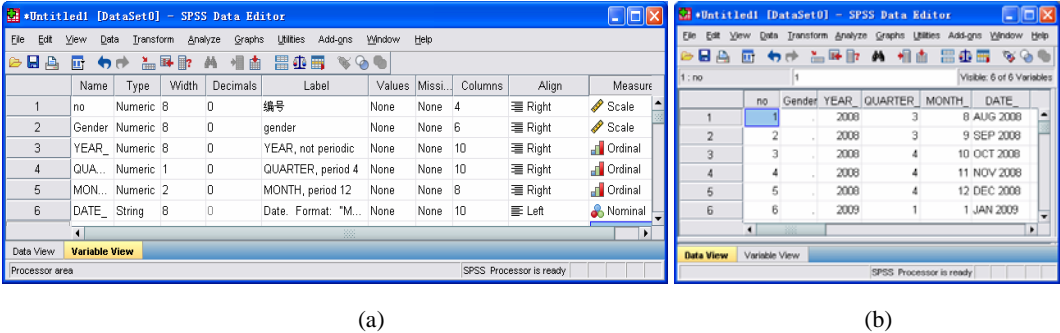


图 2-8 日期变量生成的结果（数据窗口与变量窗口）

2.1.4 数据录入与编辑

1. 录入数据

输入数据的操作方法是多种多样的，可以定义一个变量就输入这一个变量的值（纵向进行），也可以定义完所有变量后，按观测量来输入（横向进行）。

数据编辑器的二维表格中顶部标有变量名，左侧标有观测量序号。一个变量名和一个观测量序号就指定了唯一的一个单元格。可以使用上下左右箭头将插入点光标（当前单元格的定位）移动到相邻的相应位置；用 Home、End 键将插入点光标移动到同行首单元格或同行尾单元格。也可以使用滚动条或 PgUp、PgDn 上下移动一屏。

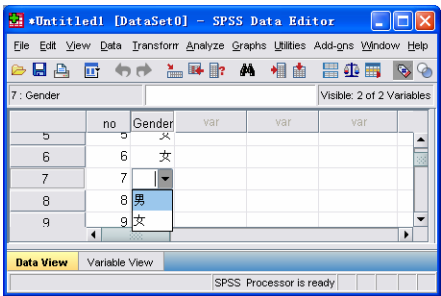


图 2-9 显示值标签的变量

按下 Value Labels 图标按钮。所有设置了值标签的变量均显示值标签，图 2-9 是显示值标签的状态下录入变量 gender 数据。当输入一个变量值时在单元格中单击向下箭头，在下拉列表选择一个定义过的值标签。

2. 编辑数据

如果知道某个变量的某个值输入错误，只要定位到相应的单元格，重新输入这个数据即可。

(1) 移动指针到指定序号的观测量

按 Edit→Go to Case 顺序单击鼠标左键，或单击工具栏上的 图标按钮，打开 Go to

对话框的 Case 选项卡, 见图 2-10(a)。在 Go to Case Number 栏输入要查找的观测量号, 例如输入 38。单击 Go 按钮。第 38 行的某个变量(取决于操作前光标停留的变量列)值被加深显示。如图 2-10(b)所示。不关闭对话框, 还可以继续查找。

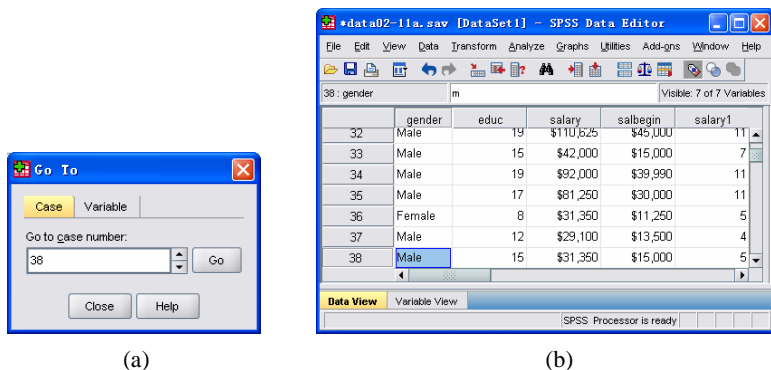


图 2-10 查找观测量对话框及查找结果

## (2) 查找变量

按 Edit→Go to Variable 顺序单击鼠标左键, 或单击工具栏上的 图标按钮, 打开 Go to 对话框的 Variable 选项卡, 单击 Go to variable 栏向下箭头, 在下拉菜单中选择要查找的变量名, 例如选择 Salbegin, 见图 2-11(a)。单击 Go 按钮。Salbegin 变量列所有值被加深显示。如图 2-11(b)所示。不关闭对话框还可以继续查找。

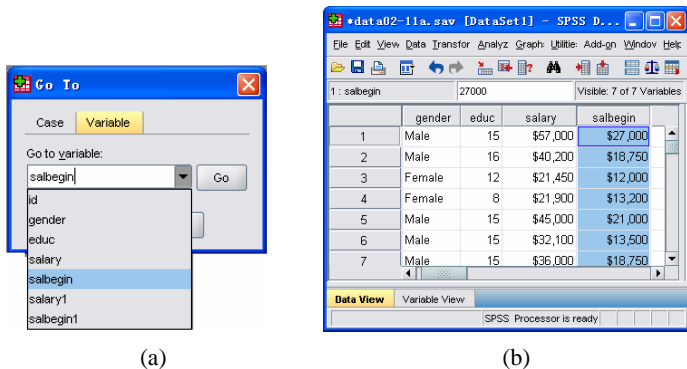


图 2-11 查找变量对话框及查找结果

Go To 对话框有两个选项卡, Case 和 Variable。可以根据查找目标进行切换。

## (3) 在 Data View 窗口中查找或替换指定变量中的指定数据(定位到单元格)

【例 2】查找 educ 值为 19 的观测量。

① 鼠标光标移至变量 educ 所在的列中任意单元格, 单击鼠标, 指定在该列中查找。

② 按 Edit→Find 顺序单击鼠标, 或单击工具栏上的 Find Data 图标按钮, 打开 Find

Data 对话框, 该对话框标题栏显示要查找的值所属变量 Column:educ, 见图 2-12。

③ 在 Find 框中输入要查找的变量数值。本例要求查找 Educ=19 的观测量, 输入 19。

④ 单击 Find Next 按钮, 即向观测量序号大的方向查找; 找到后加深显示查找内容, 见图 2-12(b)。再单击 Find Next 按钮, 可以继续查找。直到显示提示信息, 查找终止。单击 Close 按钮, 退出对话框。

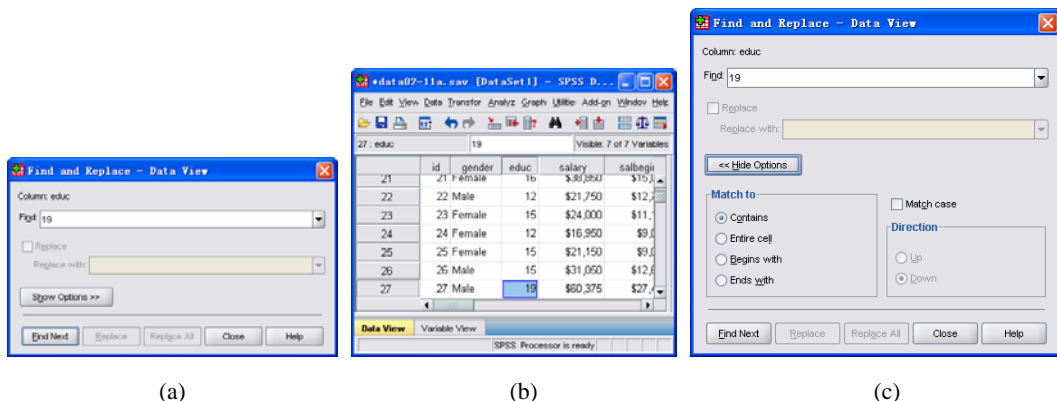


图 2-12 查找数据对话框及查找结果

#### (4) 匹配查找

单击 Show Options 按钮打开另一半窗口, 见图 2-12(c) 所示。在 Match to 栏的选项对查找进一步定义:

- Contains 包含。查找包含指定内容的变量值。例如在 Name 变量找姓张的, Find 栏填入“张”, 选择此项, 所有名字中有“张”的会一个个显示出来。
- Entire cell 整个单元格。必须整个单元格的内容完全与指定内容一致才算找到。
- Begins with 查找以指定内容开头的变量值。例如选择此项查找\$123, 找到的内容可以是\$123.0、\$1234.0, 但是不包括\$1,23.0。
- Ends with 查找以指定内容为结尾的变量值。

**注意:** 日期时间变量在 Data View 窗口中按显示格式查找。例如, 对显示格式为 10/18/2008 要查找 10-18-2008 就找不到。在 Data view 窗口中只能向观测量号大的方向查找, 不能向观测量号小的方向查找。

#### (5) 替换功能

选择 Replace, 在 Replace with 后面填写替换的内容。当查找到一个目标, 单击 Replace 按钮, 查找到的内容被替换。再单击 Find Next 按钮, 找到一个再单击 Replace 再替换一个。若单击 Replace All, 则所有与 Find 的内容匹配的都被替换成 Replace 后面的内容。

#### (6) 在 Variable View 窗口中的查找与替换

在 Variable View 窗口中, 只能对 Name、Label、Values、Missing 和自定义变量属性

列进行查找。且只能对 Label, Values 的内容和自定义的变量属性列进行替换。


**注意：**在 Values 列可以对值和值标签进行查找。但是要替换数据值就会把原来的值标签一起删除了。

#### (7) 插入一个变量

如果在现存变量的右边界左面增加一个变量，只要单击 Variable View 选项卡标签，转换到变量窗口，在变量表最下面一行，定义新变量。


如果想把要定义的变量放在已经存在的变量之间，可进行如下操作。

① 确定插入位置。在 Data View 窗口中将光标置于要插入新变量的列中任意单元格上，单击鼠标左键。或者在 Variable View 窗口中，单击新变量要占据的那行的任意位置。

② 单击 Edit→Insert Variable 命令。或单击插入变量图标按钮，在选定的位置上插入一个变量名为“Var0000n”的变量，其中“n”是系统给的变量序号。原来占据此位置的变量及其后的变量依次后移。

③ 切换到 Variable View 窗口中，对插入的变量定义属性，包括更改变量名。然后切换到 Data View 窗口输入该变量的数据。

#### (8) 插入一个观测量

观测量的排列无关紧要，其排列次序可以用排序功能整理。如果确实需要插入一个观测量，可以将光标置于要插入观测量的一行的任意单元格中，单击鼠标。单击 Edit→Insert Case 命令，或单击工具栏上插入观测量图标按钮的方法实现。结果在选中的一行上增加一个空行，可以在此行上输入该观测量的各变量值。

#### (9) 变量和观测量的删除、复制和移动

在 Data View 窗口中单击变量名；或者在 Variable View 窗口单击变量所在的行号就选择了一个变量；对变量的删除和移动可以在这两个窗口中进行。因为不允许有同名变量，因此变量不能复制。



对观测量的删除、复制和移动只能在 Data View 窗口中进行。单击一个行号就选择了这一行上的观测量。

移动变量（或观测量）只要在选择要移动的对象后，单击 Edit 菜单中 d 的 Cut 命令，找到插入位置，先插入一个空变量（或空观测量），单击空变量的变量名（或空观测量序号），即选择这个空变量（或空观测量）然后单击 Edit 菜单中的 Paste 命令，就将剪贴板中的变量（或观测量）粘贴到空变量（或空观测量）的位置上了。

要复制观测量，只要把上述步骤中的 Cut 改为单击 Edit 菜单的 Copy 命令即可。

要删除变量或观测量，只要选择要删除的对象后，单击 Delete 键或者单击 Edit 菜单中的 Clear 命令。

#### (10) 恢复删除或修改前的数据

单击Undo 图标按钮可撤销前一步操作。单击 Edit 菜单中的 Redo 命令或单击图标按钮，恢复撤销前的状态。

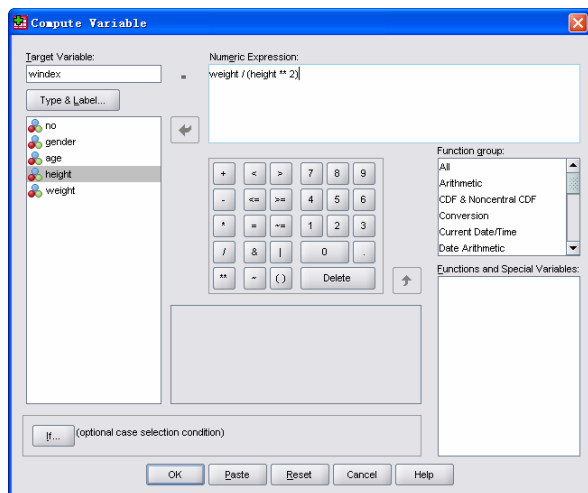
## 2.1.5 根据已有的变量建立新变量

### 1. 使用 Compute 功能完成对新变量值的计算

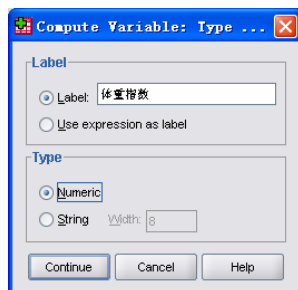
在进行数据的分析处理时往往需要根据已经存在的变量建立新变量。这一工作可以直接通过 SPSS 语句实现。对 SPSS 来说，体现其特点的更直观的方法是通过 Compute 对话框完成。

(1) 按 Transform→Compute Variable 顺序，打开如图 2-13(a)所示的 Compute Variable 对话框。

(2) 在 Target Variable 框中输入目标变量的名称，用来接收计算的值。目标变量名可以是一个新的变量名或是一个定义过的变量名。如果是新变量，单击 Type & Label 按钮，展开定义新变量类型和标签的对话框，如图 2-13(b)所示。



(a)



(b)

图 2-13 计算新变量值对话框和定义变量类型与标签对话框

① Label 栏，为新变量指定标签。

- Label，可在该框中输入长达 120 个字符的标签说明。
- Use expression as label，利用表达式的前 110 个字符作为标签。

② Type 栏，为变量指定类型。只有两种基本类型可以指定：Numeric 数值型，这是默认设置。String 字符型，要在 Width 参数框中输入字符串的宽度。

单击 Continue 按钮返回 Copmpute Variable 对话框。

(3) 在 Numeric Expression 框中组合合理的数学表达式。对话框中软键盘中包含了常数、数学运算符、关系表达符号、逻辑运算符。Numeric Expression 矩形框相当于计算器的显示屏。在数学表达式框中可以利用鼠标或键盘进行相应的编辑操作，方法如下：

① 在左面的矩形框中选择已经存在的变量，移入表达式框中；  
 ② 在操作板上选择数字或运算符，单击后出现在表达式框中；  
 ③ 在函数框中选择需要的函数，双击选中的函数；或单击选中的函数，然后单击向上箭头按钮，使选中的函数出现在表达式中。

④ 移动“1”型光标至函数名称后面的括号中，然后按①所示的方法选择自变量并单击向右箭头按钮，使其置于括号之中，代替括号中表示自变量的问号。

(4) 表达式组成规则参见 1.5.2 节的内容，另外需要注意：

① 自变量必须放在函数名后的括号中。

② 每一个关系表达式必须单独完成，例如  $age1=3$  (if  $age \geq 30$  &  $age < 40$ ) 与  $age1=4$  (if  $age \geq 40$  &  $age < 50$ ) 定义变量  $age1$  的两个值，两个值分别按变量  $age$  的不同值为条件确定，则必须分两步完成。

③ 圆点 “.” 是表达式中唯一合法的小数点符号。

(5) 条件表达式 (If)

当不同特点的观测量使用不同的表达式计算新变量的值时，新变量的值需要分步进行计算。在 Compute Variable 对话框中确定计算部分新变量值的表达式后，再利用条件表达式选择观测量。对使条件表达式值为真的观测量使用 Compute Variable 对话框中确定的表达式计算新变量的值。对那些使条件表达式为假或缺失的观测量，新变量的值或为缺失值，或保持不变。

① 在 Compute Variable 对话框中单击 If 按钮，打开 Compute Variable: If Cases 条件表达式对话框，如图 2-14 所示。

② 根据需要选择下列选项：

- Include all cases, 包括所有观测量，这是默认选项。选择此项对所有观测量使用 Compute Variable 主对话框中的计算表达式来计算新变量的值。

- Include if case satisfies condition, 只对满足条件表达式的观测量才计算新变量的值。选择此项后，激活其下面的矩形框，输入条件表达式。操作方法与 Compute Variable 中操作方法相同。

③ 条件表达式规则

大多数的条件表达式至少要包括一个关系运算符，并且可以通过关系运算符来连接多个条件表达式。例如：

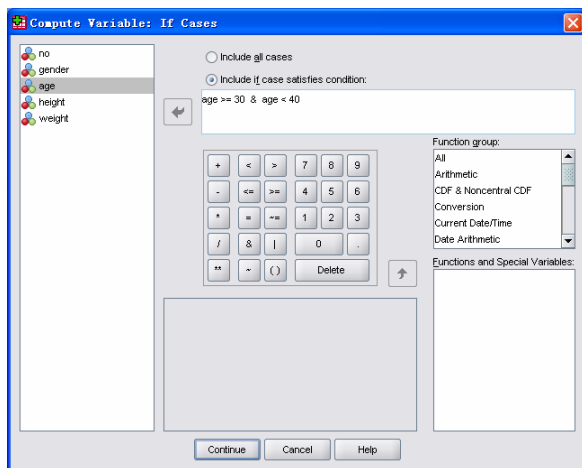


图 2-14 根据条件选择观测量子集的对话框

$\text{age} \geq 21$  表示只有 age 大于等于 21 的观测量才会被选择。

$\text{Salary} * 3 < 100000$  表示只有 Salary 乘 3 的值小于 100000 的观测量才会被选择。

$\text{Salary} * 3 < 100000 \ \& \ \text{jobcat} \neq 3$  表示只有 Salary 乘 3 小于 100000 并且 jobcat 不等于 3 的观测量才会被选择。

逻辑算符连接的两个关系表达式必须单独完成，例如  $\text{age} \geq 18 \ \& \ \text{age} < 35$  合法，而  $\text{age} \geq 18 \ \& \ < 35$  非法。

④ 单击 Continue 按钮表示确认输入的条件表达式并返回主对话框。

(6) 单击 OK 按钮，对符合 Compute Variable: If Cases 对话框中设置的条件的观测量，按主对话框中确定的计算表达式计算新变量的值。

## 2.1.6 建立值标签的工具与程序

### 1. 建立值标签的工具

为变量加标签和为变量值加标签是一件很烦琐的工作。但是对变量和变量值加了标签，并在输出表格或统计图中使用这些标签，对理解输出项的含义非常有好处。

如果分类变量的等级很多，可以先建立了变量并输入数据，然后使用加值标签工具为分类变量或标称变量加值标签。为便于说明方法，仍使用等级较少的变量举例说明该工具具有的功能和操作方法。

(1) 建立值标签工具的功能如下：

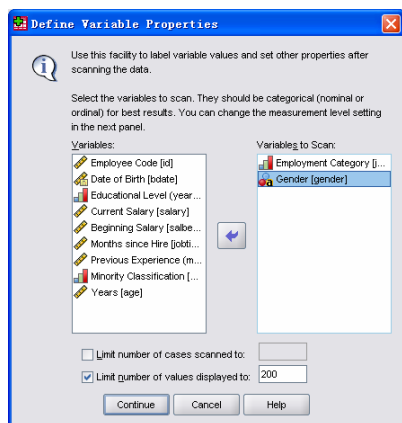


图 2-15 定义变量工具属性对话框

① 浏览查看每个变量的实际数据值并列出所有的单一取值；

② 识别未加标签的值，并提供自动标签；

③ 提供从另一个变量到所选择的变量，或从所选择的变量到多个另外的变量复制已经定义的值标签。

(2) 方法与步骤如下，使用数据 data02-01。

① 单击菜单 Data→Define Variable Properties... 打开如图 2-15 所示的对话框。左面的栏中列出了数据文件中所有的变量。

② 选择变量送入右面的栏中。选择的变量可以是：要定义值标签的变量或者是已经定义了值标签，且与要定义值标签的变量有相同的值标签；

**注意：**长字符串变量（定义长度多于 8 个字符的变量）不显示在变量表中，长字符串变量不能定义值标签和缺失值。

③ 对话框中的两个复选项：

• Limit number of cases scanned to，限制扫描的观测量数，以减少扫描时间，输入能



扫描到所有分类值的观测量数。此选项适用于大数据集。

- **Limit number of values display to**, 限制要显示的变量值的数目, 为了在错误地选择了连续变量时, 由于变量值有无限多个而增加显示长度, 输入估计的变量值的种类数。

单击 **Continue** 按钮, 进入如图 2-16 所示的主对话框。

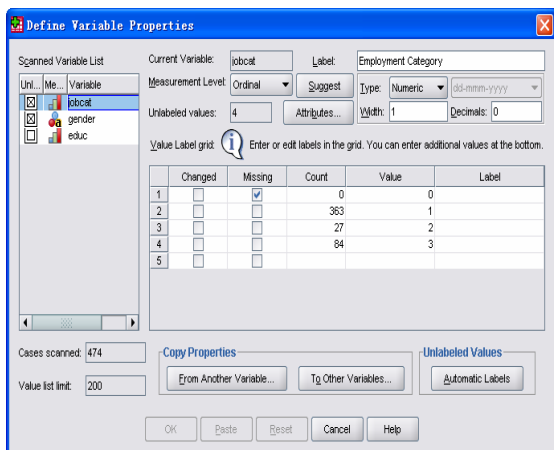


图 2-16 定义变量工具主对话框

④ 在左面的变量列表中已经定义值标签的变量前面标有□, 未经定义的变量前面标有☒。选择一个要定义的变量, 在 **Current Variable** 栏里立即显示该变量名, 其右显示该变量的变量标签; **Measurement Level** 栏里显示已经定义的测度类型; 单击向下箭头按钮, 在下拉菜单中可以选择新的定义。其右显示数值格式: 类型、宽度、小数位数。

⑤ 单击 **Suggest** 按钮打开如图 2-17 的测度类型建议对话框。上面的三栏中, 在 **Variable** 后面显示了变量名; 在 **Current Measurement Level** 后面显示了当前定义的变量测度类型, 例如图 2-17 中对 **Jobcat** 变量当前的测度类型是 **Ordinal**。读者定义的不一定正确, 因此在第 3 栏给出可以选择的两种类型, 在第 3 栏下面给出变量测度类型的建议供选择。再往下还给出了对各种测度的文字解释。据此可以在单选项中选认为正确的测度类型。

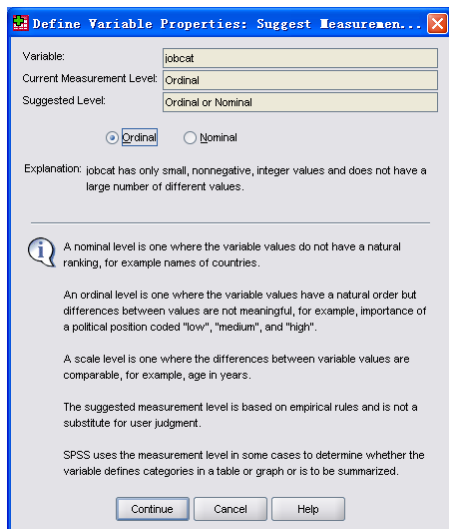


图 2-17 测度类型建议对话框



⑥ 读者可以给变量增加系统规定以外的属性。在主对话框中单击 **Attributes** 按钮打开如图 2-18 的对话框，需要增加一个属性，就单击 **Add** 按钮在 **Custom Attribute for** 栏内增加一行，在 **name** 列读者输入自定义的属性名称，在 **Value** 下面输入属性的具体内容。单击 **Continue** 按钮确认，退出对话框。

如果该变量还没有定义过各种属性，并且在变量表中有与之属性相同的变量，且值标签也可以相同，单击 **From Another Variable...** 按钮，进入相应的对话框。在对话框中

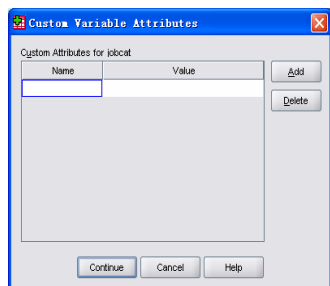


图 2-18 自定义属性对话框

选择属性相同的变量，单击 **Copy** 按钮，在主对话框中选择的变量就与这个对话框中选择的变量具有相同的属性。

**注意**，只有值相同的，值标签才相同，值不同的标签还需要再定义。

⑦ 定义值标签。主对话框下面的表中列出了所选变量的值列表，每一个值只出现一次，因此也称单值表。在 **Value** 列中列出了这些值。在 **Count** 列中，列出了相应（同行）值在数据文件中出现的频数，可以作为定义值标签的参考。

在认为是缺失值的那一行和 **Missing** 列交叉的单元格中单击鼠标左键，出现对钩，就定义了这个变量值为缺失值。

单击 **Automatic Labels** 按钮，在 **Label** 列中给出系统自动定义的值标签，一般是与值相同的字符。要给出自定义的标签，只要在 **Label** 列对应变量的单元格中输入一个值标签就可以了。输入后，在 **Change** 列中的同行单元格中出现对钩，说明已定义过了。

⑧ 进一步操作。在主对话框中，单击 **To Other Variables** 将定义好的值标签用于具有相同值标签的变量。这个变量一定是在预备对话框中已经选入变量表中的变量。

## 2. 简便操作方法

如果两个变量的值标签相同，定义好一个变量的值标签后，在 **Variable View** 窗口中，定义好的值标签的单元格中，单击右键，选择 **Copy**，就将所有值标签复制到剪贴板中。

如果另一个变量具有相同的值标签，右键单击该变量的 **Value** 列单元格，右键菜单中单击 **Paste**，就将所有剪贴板中的值标签粘贴到该行所属的变量中了。如果两个变量有不同的值，需要补充定义缺少的值标签。

## 3. 定义值标签的程序

写个小程序给一组具有相同值标签的变量加值标签，可以很方便地重复使用。例如，变量名为 *jobcat* 的只有 3 个：1、2、3，分别代表秘书、职员和经理。在数据文件中已经输入了代码 1、2、3。现在要加上值标签，程序如下：

```
Value labels jobcat
```

```
1 '职员'
```

2 '秘书'

3 '经理'.

**注意：**Value labels 是语句关键字不能改变，其后的变量可以是空格隔开的变量表。

变量值在前，标签在后，标签字符串加单引号。作为值标签的字符串中没有特殊字符，如“/”时，也可以不加引号。


最后以圆点结束。

单击“运行”按钮，执行程序，值标签加在工作数据文件的相应变量上。

在 Data 菜单中还有可以使用外部 SPSS 数据文件作为模板文件的功能，为工作数据文件定义文件属性、变量属性，诸如变量类型、测度类型、变量标签、值标签、显示格式（宽度、小数位数等）、写和打印格式，甚至变量集、多响应变量集等。该功能菜单在 Data 菜单的第二项，Copy Data Properties...。由于在实际工作中，如此相同的数据文件不多，有以上方法操作已经够用，不再赘述。

## 2.1.7 打开、保存与查看数据文件

### 1. 打开一个已有的数据文件

按 File→Open 顺序单击鼠标，或单击工具栏上的图标按钮，打开 Open File 对话框。在“搜索”框中指定文件保存位置。数据文件类型为\*.sav。找到或输入要打开的数据文件，双击之，就可以将数据文件显示在数据窗口中。

在打开文件对话框中，单击“文件类型”框内向下箭头，展开 SPSS 允许打开的文件类型表列。数据文件的类型大致有以下几种。

**SPSS (\*.sav):** SPSS 建立的数据文件，扩展名为“\*.sav”。

**SPSS/PC+ (\*.sys):** SPSS/PC 或 SPSS/PC plus 建立的语句文件，扩展名为“\*.sys”。

**SYSTAT (\*.syd、\*.sys):** SYSTAT 建立的数据文件扩展名为“\*.syd”或语句文件，扩展名为“\*.sys”。

**SPSS portable (\*.por):** 用 SPSS 简便格式保存的数据文件。

**Excel (\*.xls):** Excel 建立的表格数据文件。SPSS 可以直接打开 Excel 电子表格文件。

**Lotus (\*.w\*):** 用 Lotus 1-2-3 格式写的文件。可以是 1A 版、2 版、3 版 Lotus1-2-3 记录的数据文件。它的一行转换成一个观测量，变量是一列。

**Sylk (\*.slk):** 用 Sylk 格式保存的数据文件。

**dBASE (\*.dbf):** 数据库格式文件，扩展名为“\*.dbf”。可以是各种版本 dBASE 或 FoxBase 建立的数据库文件。一个记录转换成数据窗口中的一个观测量。

**SAS (\*.sas7bdat,\*.sd7,\*.sd2,\*.ssd01,\*.xpt)**各版本的 SAS 软件生成的数据文件。

**Stata(\*.dta):** Sdata 软件生成的数据文件。

**Text (\*.txt,\*.dat):** 纯文本数据文件和用 ASCII 码编写的文件。

## 2. 保存数据文件

保存数据文件可以使用 **File** 菜单中的 **Save** 和 **Save As** 命令。操作方法与 **Windows** 系列应用软件的文件保存方法一样，而 **SPSS** 数据文件可以选择不同变量保存为不同的文件。

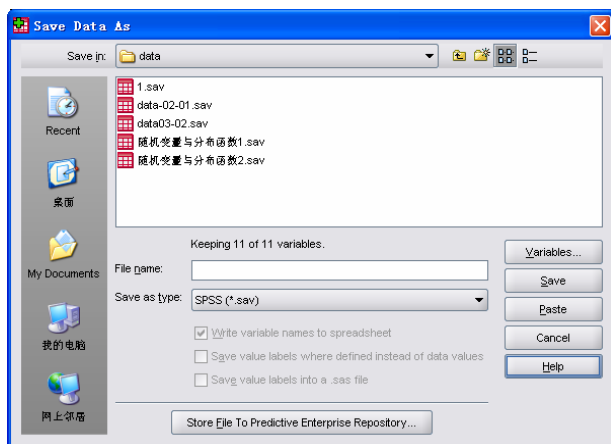
可选择的数据文件的类型很多。基本上可以打开的文件类型，都是可以保存的文件类型，但大部分会丢失变量标签和值标签。保存为文本文件时有下面两种类型。

**Tab-delimited (\*.dat)**: 保存为 **ASCII** 码文件，用制表符作为两个观测量之间的分隔符。如果一个软件不能读取其他任何格式的数据文件，可以使用此种格式保存数据。在将数据保存为此种格式文件的同时，变量标签、值标签、缺失值定义均丢失。

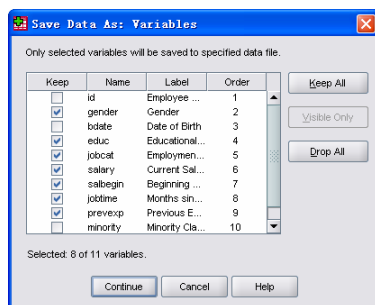
**Fixed ASCII (\*.dat)**: 保存为固定列格式的 **ASCII** 码文件。

## 3. 保存部分变量

单击 **File**→**Save As**，打开如图 2-19(a) 所示的 **Save Data As** 对话框，在“保存在”栏中设置保存位置，在“文件名”栏中输入文件名。



(a)



(b)

图 2-19 保存数据文件及保存子集文件

单击“保存”对话框中的 **Variables** 按钮，打开如图 2-19(b)所示的 **Save Data As: Variables** 对话框。在该对话框中选择要保存的变量。系统默认 **Keep All**，所有变量名前都标有对钩。只有标有对钩的变量才被保存到文件中。选择要保存的变量或单击 **Keep All** 全部保存，去掉不保存的变量；或者选择 **Drop All** 全部不保存，选择要保存的变量。单击 **Continue** 按钮，返回 **Save As** 主对话框，单击 **Save** 按钮。完成保存部分变量的操作。

**【例 3】** 另存为 **ASCII** 码数据文件实例。

数据文件 **data02-02.sav** 中有 5 个变量：**no**（编号）、**gender**（性别）、**age**（年龄）、**height**（身高）、**weight**（体重）。把数据保存为固定格式的 **ASCII** 码文件的操作如下：

(1) 按 File→Save As 顺序单击鼠标左键, 展开 Save As 对话框;

(2) 在 Save As 对话框中指定存储位置(驱动器、目录), 选择文件类型为 Fixed ASCII, 其扩展名为\*.dat。并输入文件名保存。在输出窗中显示保存记录如表 2-1。

表 2-1 中第一列是变量名。第二列是该变量所在的记录号, 这些变量处于同一个记录中。第三列是对应的变量所占的起始列号, 第四列是对应的变量所占的结束列号, 变量 *no* 占 2 列, 变量 *gender* 占 2 列……第五列是对应的变量的格式。前两个变量是字符型, 后三个变量是数值型。由于各变量值间没有空格, 如果在其他软件中打开此文件, 该表对重新整理数据很重要, 应该保存。

表 2-1 文件以 ASCII 码形式存入指定位置

Variable	Rec	Start	End	Format
no	1	1	2	F2.0
gender	1	3	4	F2.1
age	1	5	7	F3.2
height	1	8	12	F5.2
weight	1	13	15	F3.2

当前数据编辑器中定义的所有变量名和测度类型图标。鼠标单击变量表列中一个变量, 右半部分 Variable Information 变量信息显示区, 列出指定变量的属性。Variable Information 框中只能显示一个变量的属性信息。例如, 图 2-20 中显示的是职别 jobcat 变量的信息: 第一行是变量名, jobcat; 第二行是变量标签, Label: Employment Category; 第三行是变量类型, Type: F1, 表示 1 位的数值型变量; 第四行是变量的缺失值定义, Missing Value: 0, 说明 0 为缺失值。空行后是值标签 Value Labels: 值 1 标签为办事员; 值 2 标签为保管员; 值 3 标签为经理。


单击 Close 按钮, 关闭变量对话框, 返回到数据窗口。

#### 5. 查看文件信息

可以利用 File 菜单的命令查看所有定义的变量。方法是鼠标按 File→Display Data File Information 顺序单击菜单项, 在二级菜单中:

(1) 单击 Working File, 当前数据窗口中所有变量的有关信息显示在输出窗 Variable Information 表中;

#### 4. 查看变量信息

在数据窗口中选择一个变量, 单击 Utilities 菜单中 Variables 命令, 打开 Variables 变量信息对话框, 如图 2-20 所示; 也可以单击  Variable 图标按钮, 打开该对话框。

对话框中左半部是变量列表, 列出

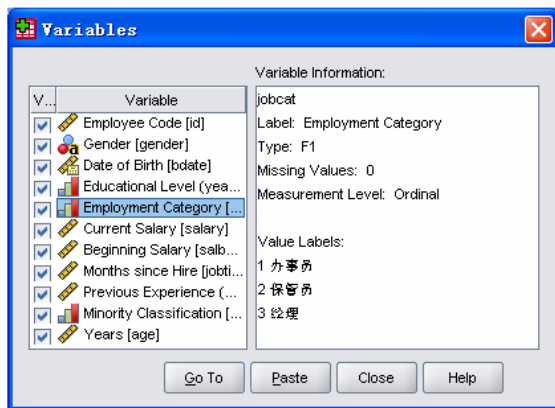


图 2-20 变量信息对话框

(2) 单击 External File，打开 Display Data info 对话框，指定一个外部数据文件，文件中所有变量信息显示在输出窗口中。

文件信息包括文件保存位置、文件类型、生成日期以及是否定义了加权变量等。变量信息包括变量在数据编辑窗口中的位置序号、变量名、变量标签、值标签、格式和缺失值。

## 2.2 数据文件的转换

### 2.2.1 ASCII码数据文件的转换

几乎所有有计算功能或管理数据功能的软件，都可以输出 ASCII 码数据文件。因此掌握 ASCII 码数据文件转换成 SPSS 数据文件的方法是非常重要的。

#### 1. 不同格式的 ASCII 码数据文件

SPSS 可以读入 ASCII 码数据文件并将其转换为 SPSS 格式，显示在数据窗口中。ASCII 码数据文件有固定宽度格式（Fixed width）和使用分隔符的自由格式（delimited）两种。所谓固定格式即一个观测量（或称记录）占一行或若干行，每个变量所占起始列和结束列是固定的，如图 2-21 所示。自由格式即每个变量在文件中的列位置不一定是固定的，各变量值之间使用相同的符号（如空格或逗号）隔开，转换时根据分隔符和变量值排列顺序进行。

(1) 固定格式排列的 ASCII 码数据文件中，数据的排列方式有以下两种。

① 每行安排一个观测量，每个变量值之间由空格分隔，见图 2-21(a)。这种固定格式也可以看作使用分隔符的自由格式数据文件。见 data02-02a.txt。

② 每行安排一个观测量，但变量值之间没有任何分隔，如图 2-21(b)所示，数据安排实例见 data02-02b.txt。

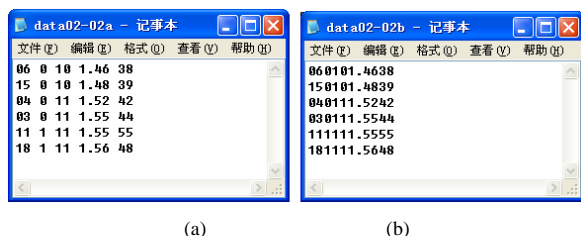


图 2-21 不同排列的固定格式 ASCII 码数据文件

#### (2) 使用分隔符的自由格式 ASCII 码数据文件

① 每行安排若干个观测量，如图 2-22(a)。或整齐地排列两个观测量，如图 2-22(b)所示。看上去既是固定格式 ASCII 码数据文件，又是使用分隔符的自由格式 ASCII 码数据文件。转换程序将其归为使用

分隔符的自由格式文件。如果使用固定格式转换的操作，可以在转换后通过对数据文件的编辑才能形成正确的转换结果。

② 每行一个或多个观测量，使用分隔符将各变量值分开，甚至一个观测量从一行中间开始并在下一行继续，如图 2-22(b)所示，见数据文件 data02-04.txt。

beername	calorie	sodium	alcohol	cost
Budweiser	144	19	4.7	.43
Innenbrau	157	15	4.9	.48
Heineken	152	11	5	.77
Buccherger	175	24	5.5	.4
Miller-lite	99	18	4.3	.43
Coors	140	16	4.6	.44
Michels-lich	125	11	4.2	.5
Heirln	149	6	5	.79
Hanns	136	19	4.4	.43
Olympia-gold	72	6	2.9	.46

(a)

beername	calorie	sodium	alcohol	cost
Budweiser	144	19	4.7	.43
Innenbrau	157	15	4.9	.48
Heineken	152	11	5	.77
Buccherger	175	24	5.5	.4
Miller-lite	99	18	4.3	.43
Coors	140	16	4.6	.44
Michels-lich	125	11	4.2	.5
Heirln	149	6	5	.79
Hanns	136	19	4.4	.43
Olympia-gold	72	6	2.9	.46

(b)

图 2-22 不同排列的自由格式 ASCII 码数据文件

## 2. 固定格式 ASCII 码数据文件的转换

以图 2-21(b)为例说明固定格式 ASCII 码数据文件转换为 SPSS 数据文件的操作。

图中所示数据文件为 data02-02b.txt，数据由 5 个变量组成，变量编号占 1、2 列，性别占第 3 列，年龄占第 4、5 列，身高占 6~9 列（小数点占一列），体重占 10、11 列。

数据文件转换步骤如下：

(1) 按 File→Read Text Data 顺序展开 Open File 对话框，指定一个扩展名为 txt 的数据文件（data02-02b.txt）并单击“打开”按钮，展开 Text Import Wizard 对话框，如图 2-23 所示。分 6 步完成转换工作，此为第 1 步。数据显示在下面的带有标尺的预览框内。

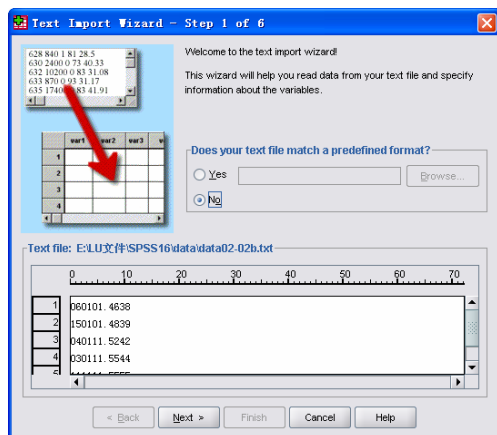


图 2-23 第 1 步：Text Import Wizard 对话框

右面的 Does your text file match a predefined format 栏询问文本文件是否要与事先定义的格式匹配，如果需要，单击 Yes 并通过单击 Browse 按钮指定一个扩展名为 tpf 的文件。通常不选择该选项。单击 Next 按钮，打开 Step 2 of 6 第 2 步的对话框，如图 2-24 所示。

(2) 在第 2 步对话框中回答两个问题：

① How are your variable arranged 栏，如何安排你的变量？选择 Delimite 是使用分隔符将变量隔开的；选择 Fixed Width 是使用固定列宽度。图 2-24 预览框中的数据排列整齐，列宽是固定的，因此选择 Fixed Width 选项。

② Are variable names included at the top of your file 栏，数据文件顶部是否包括变量



名? yes 或 no 选其一。移动滚动条可以看到，顶部没有变量名，因此选择 no。单击 Next 按钮打开第三步对话框，如图 2-25 所示。

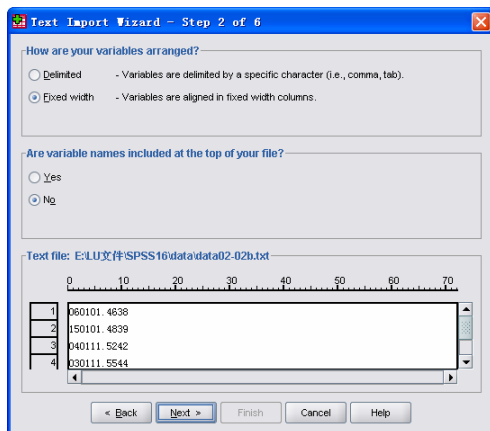


图 2-24 第 2 步：指定数据排列方式

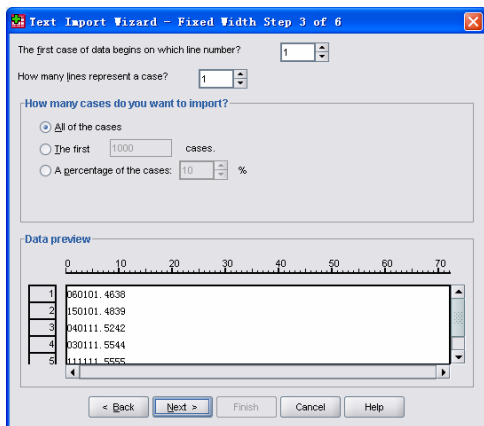


图 2-25 第 3 步：设置参数

(3) 在第 3 步对话框中要求提供有关观测量的信息。一个观测测量相当于数据库中的一个记录。各选项的含义与操作如下：

① The first case of data begins on which line number 参数框，要求指定数据文件中第一个包括数据值的行号，默认值为 1。如果在顶行包括了对变量的解释文字或变量标签，该值不能是 1。应该在该项后面的数值栏中设置具体值。

② How many lines represent a case 参数框，要求回答一个观测测量占几行，以确定何处为一个观测测量的结束位置和下一个观测测量的起始位置。本例每个观测测量占一行。

③ How many cases do you want to import 栏，指定想转换的观测测量数。

- All of the cases，指定转换所有观测测量。此为默认的选择。
- The first  cases，指定前  $n$  个观测测量， $n$  是自定义的正整数。框中输入  $n$  值。
- A percentage of the cases，指定一个百分数，转换系统按指定的百分比随机提取观测测量。由于随机采样是通过每个观测测量产生一个独立的伪随机数进行的，因此该百分比是一个近似值，最后采样得到的样本占观测测量总数的百分比接近这个指定值。本例指定转换所有观测测量，选择第一项。

(4) 第 4 步窗口画面如图 2-26 所示。在浏览窗口中，加竖线将各变量值分开，标明将如何读取数据。浏览区上有尺，左有观测测量号。插入分隔线和去除分隔线的方法有二：

① 需要插入变量分隔线处单击鼠标键，出现分隔线。本例的 ASCII 码数据文件中各变量值间没有分隔符，因为 1、2 列为编号值，第 3 列为性别值，因此要在第 2、3 列之间加分隔线。在列间单击鼠标左键，就会插入一根分隔线；右键单击一根已经存在的

分隔线, 该分隔线变成蓝色, 再单击 Delete Break 按钮, 则删除该分隔线。

② Column Number, 在第  $n$  列右需要加分隔线则输入该列的列号  $n$ 。单击 Insert Break 按钮, 则在第  $n$  列右侧加入分隔线, 单击 Delete Break 按钮, 将第  $n$  列右侧分隔线删除。

由计算机产生的连续数据流, 各变量值之间没有空格或其他分隔符, 很难确定一个观测量从哪里开始, 到哪里结束, 应该使用其他应用程序对其重新编辑成便于转换的排列。

(5) 第5步对话框如图2-27所示。这一步确定变量名和变量类型。对话框预览栏内显示出根据第4步的变量分隔线划分的各变量数据。变量名为系统默认的  $Vn$ 。 $n$  为自左至右的变量顺序号。转换程序据此对各变量读取并转换形成 SPSS 数据文件。

① 在预览栏中单击要定义的默认变量名。在 Variabl name 下面输入自己命名的变量名。除应该复合变量名的有关规定外, 不能重名。

② Data format 下拉列表中选择一种类型, 定义选中变量的数据类型。

(6) 第6步对话框如图2-28所示。

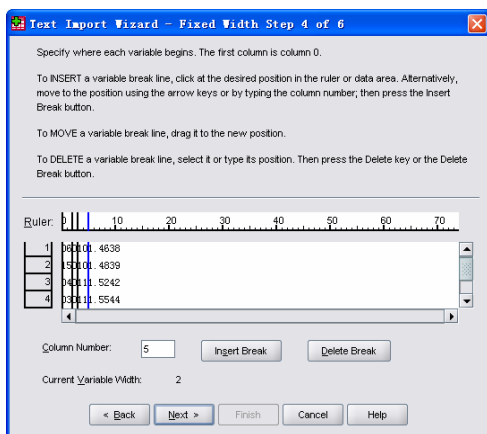


图 2-26 第4步: 变量间加分隔线

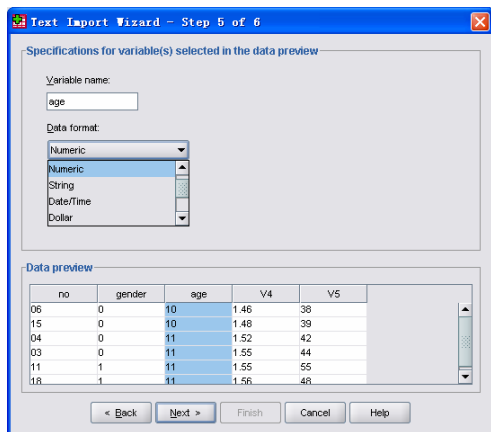


图 2-27 第5步: 定义变量名和类型

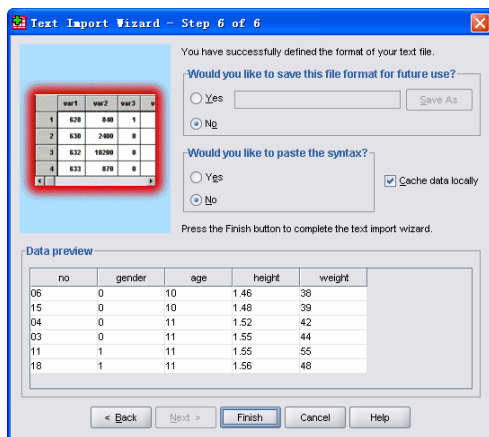


图 2-28 第6步: 最后一步保存格式

① 在 Would you like to save this file format for future use 栏内选择 Yes, 指定将该格式保存到一个文件中, 以便对相同或类似数据文件进行转换时使用。单击 Save as 按钮, 打开相应的对话框, 指定保存位置和文件名。否则单击 No 按钮。



② Would you like to paste the syntax 栏内选择 Yes, 把各步确定的转换参数粘贴到语句窗口形成命令文件, 以便进行类似转换工作时使用。否则单击 No 按钮。

一切参数设置工作完成后, 单击 Finish 完成按钮, 转换开始, 见图 2-28。最后在数据编辑窗口中显示转换结果, 见图 2-29。在数据编辑窗口中对各变量的标签、值标签、缺失值等属性再进行完善和修改。

The figure consists of two side-by-side screenshots of the SPSS Data Editor window. The left window is titled 'Untitled2 [DataSet2] - SPSS Data Editor' and is in 'Variable View'. It shows a list of variables: 'no', 'gender', 'age', 'height', and 'weight'. Each variable has a 'Name', 'Type' (Numeric), 'Width', and 'Decimals' specified. The right window is also titled 'Untitled2 [DataSet2] - SPSS Data Editor' but is in 'Data View'. It shows a table with 6 rows of data. The columns are 'no', 'gender', 'age', 'height', and 'weight'. The data is as follows:

	no	gender	age	height	weight
1	6	0	10	1.46	38
2	15	0	10	1.48	39
3	4	0	11	1.52	42
4	3	0	11	1.55	44
5	11	1	11	1.55	55
6	18	1	11	1.56	48

图 2-29 在数据编辑的两个窗口中的转换结果

### 3. 自由格式 ASCII 码数据的转换

自由格式 ASCII 码数据的文件有以下特性:

① 各观测量中的各变量值按相同顺序排列, 但同一变量的值不一定占有相同的列位置。

② 两个值之间以空格、逗号或其他符号分隔。

③ 每行可以有不止一个记录 (一个记录即一个观测量)。

在转换时, 读完最后一个定义的变量的值就读完了数据文件中的一个观测量, 然后 SPSS 读下一个值时就认为是下一个观测的第一个变量的值。因此, 定义的变量数必须与 ASCII 码数据文件中的变量数目相同, 否则转换后的结果是混乱的。当存在两种类型的变量时, 会出现数据与变量类型不匹配的错误。

【例 4】data02-04.txt 是一组 12 盎司啤酒中的成分和价格的 ASCII 码数据文件。如图 2-22(b)所示。是一个空格分隔、一行两个记录的 Text 文件。变量包括 beername (啤酒名)、calorie (热量卡路里)、sodium (钠含量)、alcohol (酒精含量)、cost (价格) 共 5 个变量。空格做分隔符, 且空格数不定。

转换为 SPSS 格式数据文件的操作步骤是:

(1) 按 File→Read Text Data 顺序展开 Open File 对话框, 指定一个扩展名为 txt 的数据文件 data02-04.txt 并单击“打开”按钮, 展开 Text Import Wizard 对话框, 如图 2-30 所示。分 6 步完成转换工作, 此为第 1 步。数据显示在预览框内。可以看出数据间空格做分隔符, 每行一个记录但排列较乱。

右面的 Does your text file match a predefined format? (询问文本文件是否与事先定义

的格式匹配) 栏, 如果是, 单击 **Yes** 按钮并通过单击 **Browse** 按钮指定一个扩展名为 .tpf 的文件。通常不选择该选项。单击 **Next** 按钮, 打开 Step 2 of 6 第 2 步的对话框, 如图 2-31 所示。

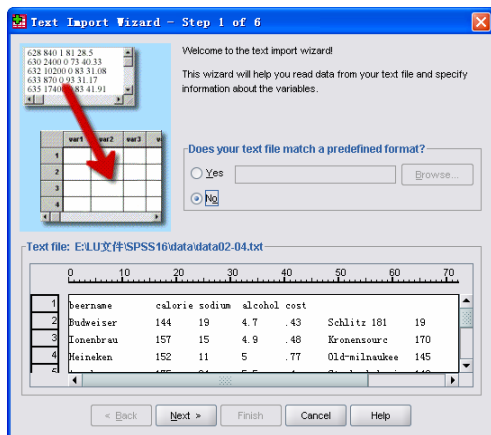


图 2-30 第 1 步: 打开 ASCII 码数据文件

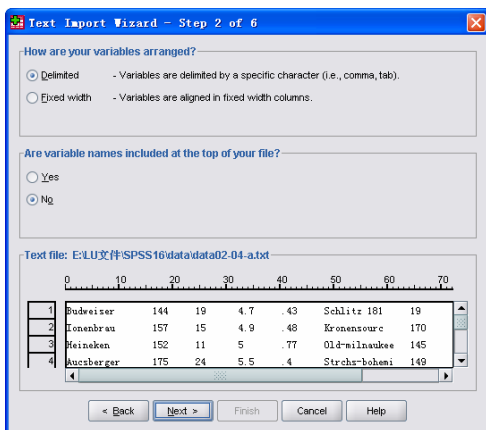


图 2-31 第 2 步: 指定数据排列方式

(2) 在第 2 步对话框中回答两个问题:

① How are your variables arranged 栏, 如何安排变量有两个选项: **Delimited**, 用分隔符隔开变量值; **Fixed Width**, 固定列宽度。

图 2-31 中的预览框中的数据排列凌乱, 不是固定列宽度, 但每两个数据之间均有空格, 是使用分隔符的, 因此选择 **Delimited**。

② Are variable names included at the top of your file 栏, 数据文件顶部是否包括变量名? **Yes** 或 **No** 选其一。移动滚动条可以看到, 顶部没有变量名, 因此选择 **No**。单击 **Next** 按钮打开第 3 步对话框如图 2-32 所示。

(3) 在第 3 步对话框中提供有关观测量的信息。

① The first cases of data begins on which line number 参数框, 指定数据文件中第一个包括数据值的行号, 默认值为 1。本例顶行没有变量名, 所以数据从第 1 行开始。参数框内数值应为 1。

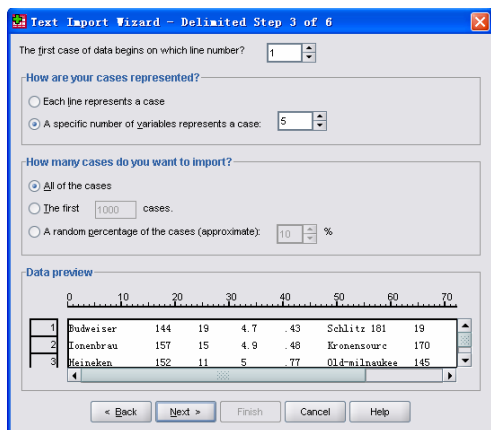


图 2-32 第 3 步: 参数设置

② How are your cases represented 栏, 你的观测是怎样描述的, 实际上是问一个观

测量占几行? 以便确定何处为一个观测量的结束位置和下一个观测量的起始位置。本例为两个观测量占一行。此项不能确切回答数据文件排列的实际情况。其中有两个选项:

- **Each line represents a case**, 每行仅包括一个观测量, 即使变量非常多, 使得一行非常长, 也属于这种情况。如果各行包括的数据量不同, 每个观测量包括的变量数由数值最多的行确定。数值较少的观测量, 多出来的变量值赋予缺失值。

- **A specific number of variables represents a case**, 指定每个观测量包括的变量数, 告诉系统在何处停止读一个观测量, 并开始读下一个观测量。该选项允许在同一行中有多个观测量或一个观测量开始于一行的中部, 并在下一行继续。系统根据数值个数读取数据, 不管行数。因此每个观测量必须包括所有变量的数值 (缺失值必须使用分隔符指定), 才能正确进行转换。本例一个观测量包括 5 个变量因此选择此项, 设置数值 5。

③ **How many cases do you want to import** 栏, 指定想转换的观测量数。

- **All of the cases**, 指定转换所有观测量。此为默认的选择。

- **The first ☐ cases**, 指定前  $n$  个观测量。 $n$  是由读者输入的正整数。

- **A random percentage of the cases (approximate)**, 指定一个百分数, 转换系统按指定的百分比随机提取观测量。由于随机采样是通过对每个观测量产生一个独立的伪随机数进行的, 因此该百分比是一个近似值。本例指定转换所有观测量, 选择第一项。

单击 **Next** 按钮, 打开如图 2-33 所示的第 4 步对话框。

(4) 第 4 步指定分隔符和字符串的标识符。

① **Which delimiters between variables** 变量间分隔符设置。

栏中列出的分隔符有 5 种, **Tab** (跳格)、**Space** (空格)、**Comma** (逗号)、**Semicolon** (分号)、**Other** (其他)。可以同时选择几种。还可以选择 **Other** 项, 并在其后的文本框中输入一个分隔符。根据指定的分隔符, 转换后的数据文件状态显示在预览栏中, 可以查看所指定的分隔符是否有误。本例数据中有缺失值, 所以选择 **Tab**, 预览窗口中的观察显示结果选择 **Tab** 效果最佳。

② **What is the text qualifier** 文本限定标志设置。下设 5 个选项: **None** (没有限定符)、**Single quote** (单引号)、**Double quote** (双引号)、**Other** (其他), 选择 **Other** 需要在其后面框中输入一个具体的限定标准。本例中, 字符串没有加单引号或双引号标识, 所以在右栏中选择 **None**。

单击 **Next** 按钮, 打开如图 2-34 所示对话框。

(5) 第 5 步定义每个变量值的变量名和数据格式, 以便在进行转换并组成数据文件时读取各变量值。在预览栏内选择一个变量, 对它进行定义。选择时单击要定义的一列数据顶部的默认变量名, 默认变量名出现在 **Variable name** 栏内。

① **Variable name**, 删掉或覆盖默认的变量名, 输入自己定义的变量名。

② **Data format**, 在下拉列表中选择变量类型。

本例定义变量 **beername** 为字符型变量, 对字符型变量还要在后面的 **Characters** 栏中

输入字符串长度。本例输入最长的字符数 13；本例还定义变量 *calorie*、*sodium*、*alcohol*、*cost* 为数值型。

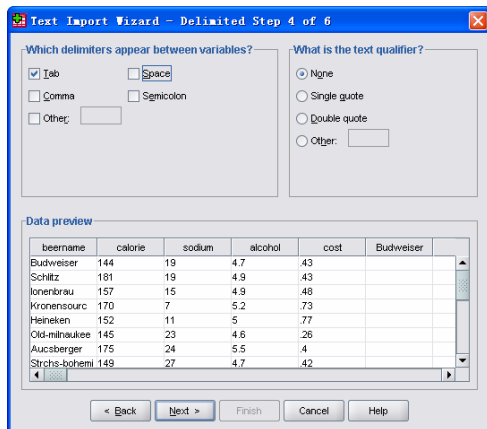


图 2-33 第 4 步：指定分隔符图

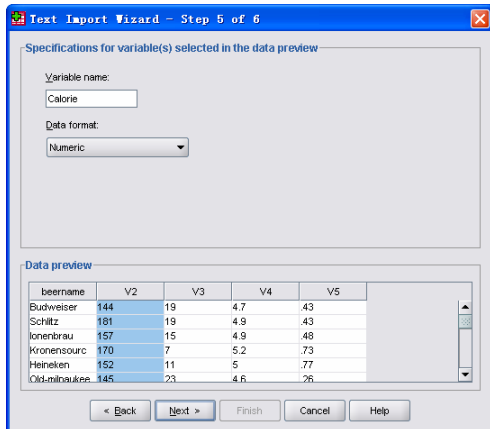


图 2-34 第 5 步：定义变量类型等属性

(6) 最后一步如图 2-28 所示，只需回答两个问题。

① 是否要保存转换使用的数据格式以便以后使用。如果需要，则单击 **Save As** 按钮指定存储位置和文件名。本例回答 No。

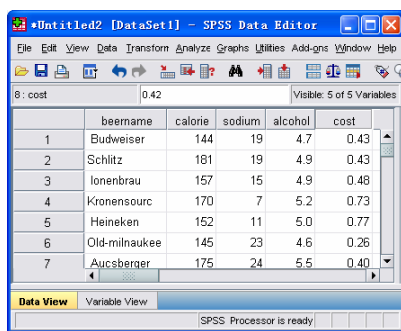
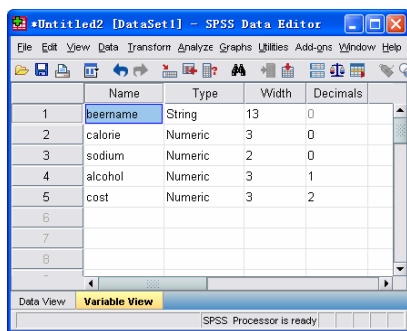


图 2-35 数据编辑窗口中的转换结果

② 是否要将其转换为 SPSS 命令语句。

选择过后，单击 **Finish** 按钮，系统开始进行转换。转换后的数据出现在数据编辑窗口中，如图 2-35 所示。在数据编辑窗口中对转换后的数据进行编辑。例如调整每个变量所占宽度等。

## 2.2.2 数据库文件的转换

任何数据库文件，例如 Excel、dBase、FoxBase、FoxPro、Oracle，要使用 SPSS 软

件进行分析处理,就必须将数据库文件转换为 SPSS 格式。

### 1. 快速完全转换

快速完全转换就是打开对话框选择一种数据库文件直接打开。以打开中国女排档案的 Excel 文件为例。

(1) 按 File→Open→Data 顺序打开 Open File 对话框,建立搜索路径。

(2) 展开文件类型菜单,选择 Excel 类型,选择中国女排档案文件 data02-17.xls。

(3) 单击 Open 按钮,打开 Opening Excel Data Source 对话框,见图 2-36。在对话框中设置:

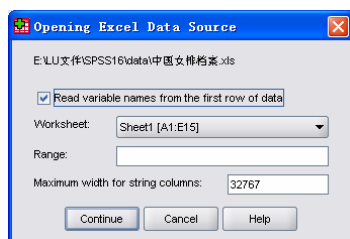


图 2-36 指定转换参数的对话框

① Read variable names from the first row of data 是否从数据文件第 1 行读取变量名。对 Excel 文件来说,回答是肯定的。因此选择此项。

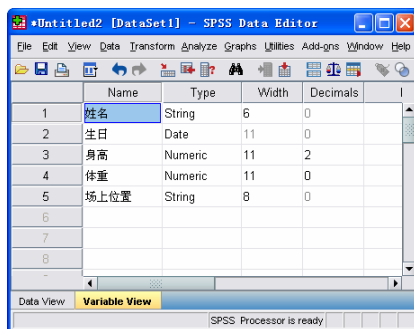
② Worksheet 工作簿。在这项中指定工作簿中的工作表和读取数据的范围。默认值是系统对所指定的 Excel 文档的分析得出的,工作表是 sheet1,数据范围是 A1:E15。

③ Range 部分无须再指定范围。

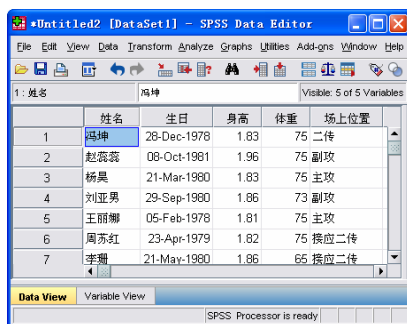
④ Maximum width for string columns 最大的字

符串列宽度。

单击 Continue 按钮,转换自动进行。结果见图 2-37。



(a)



(b)

图 2-37 转换结果

### 2. 选择部分数据库字段转换为 SPSS 数据文件

SPSS 16.0 虽然可以直接打开许多类型的数据文件,例如直接打开 Excel 数据文件,但是,转换成的 SPSS 数据文件包括所有变量,如果要选择某些变量,或者数据库中的字段组成 SPSS 数据文件,使用下述方法更方便。

(1) 按 File→Open Database→New Query 顺序展开 Database Wizard 对话框,如图 2-38

所示。

右栏中显示的是一些已经安装了原软件的数据库类型。如果想转换的数据文件是图 2-38 主对话框中不包括的类型，例如 Lotus 1-2-3、Visual FoxPro 数据库等，可以单击 Add ODBC Data Source 按钮打开下一个对话框，增加新的数据源。但是必须事先安装了相应的软件，以便使用其数据库驱动程序，否则无法完成添加数据库类型的工作。

(2) 在对话框中的数据来源栏内选择一种数据源，如选择 Excel 数据文件。

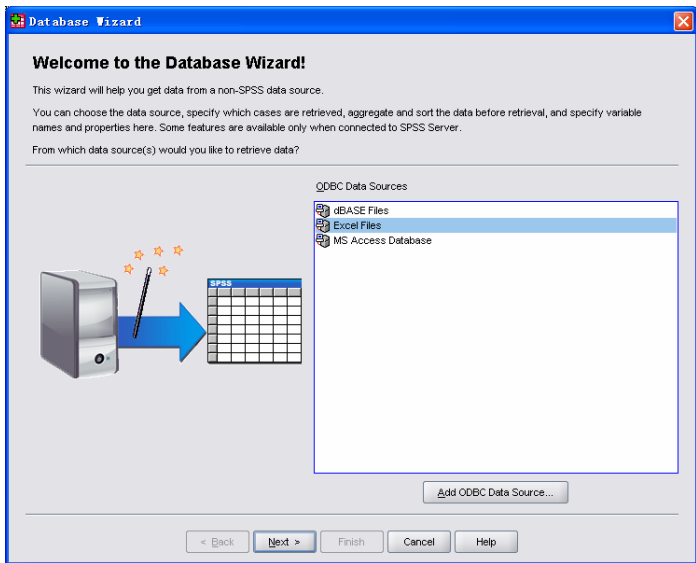


图 2-38 数据获取工具主对话框

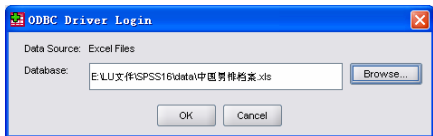


图 2-39 选择具体数据库文件

数据库文件，在下一个窗口中列出各字段，见图 2-40。

(3) 在 Available Tables 栏选择变量或字段。

- 如果指定的 Excel 文件，在左栏中显示所有的工作表，单击工作表前面的十字图标，显示变量表。
- 如果指定的是数据库文件，单击数据库文件名前面的十字图标，直接在左窗口显示字段列表。
- 可以单击另一个工作表 Sheet（或另一个数据库）选择需要的变量（或字段）；打开两个以上工作表（或数据库），选择需要变量（或字段）。

(4) 鼠标单击 Excel 工作表中的变量名（或数据库的字段名），单击向右箭头按钮，将显示到右栏中，见图 2-40。

(5) 排列变量出现顺序。选择完成后，还可以在右栏 Retrieve Fields in This Order 中选择一个变量后，单击上下箭头按钮改变变量的排列顺序。单击 Finish 按钮，按右窗口

• 单击 Next 按钮。在如图 2-39 所示对话框中，单击 Browse 按钮，指定一个想要进行转换的 Excel 文件。单击 OK 按钮后，在下一个窗口列出 Excel 工作簿中所有工作表。

• 如果选择了数据库文件，例如 DBF 文件，单击 Next 按钮，系统将找到指定类型的数

中变量排列顺序转换。在数据编辑窗口中显示转换结果。

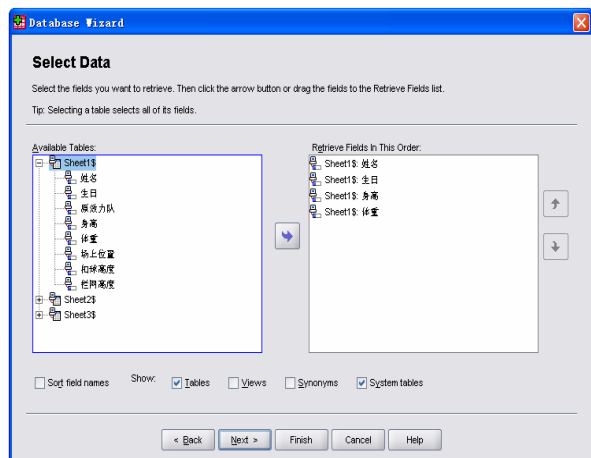


图 2-40 选择转换为 SPSS 变量的字段

(6) 选择观测量。如果需要根据表达式选择数据库中的某些观测量,可以在变量选择完成后,单击 **Next** 按钮,进入如图 2-41 的对话框,按给定的表达式、指定的方法和给定的数量选择观测量。

① 在 **Field** 栏选择变量(或字段)双击之,变量出现在 **Criteria** 栏的第一个 **Expression1** 列中,单击向下箭头按钮,在下拉菜单中可以选择其他变量(或字段);在 **Relation** 列的下拉菜单中选择关系运算符;在第二个

**Expression** 列给出关系表达式的另一部分。这样,在第一行的 **Expression1**、**relation**、**Exepretion2** 三者形成一个完整的表达式,作为选择观测量的标准。

② 选择观测量的方法。在 **Sampling Method** 栏中有两个选项:

- **Retrieve cases and randomly select in SPSS**, 先转换观测量,在 SPSS 中随机选择。

- **Randomly select in database and retrieve into SPSS**, 在数据库中随机选择,再转换到 SPSS 里。

如果先转换然后在 SPSS 中选择,不能获得数据库的子集。

③ 采样数量的确定。在 **Sample size** 栏中有两个选项:

- **Approximate percentage of all cases** 后面给出百分数。系统按这个百分比确定采样数。这个数字是近似的。

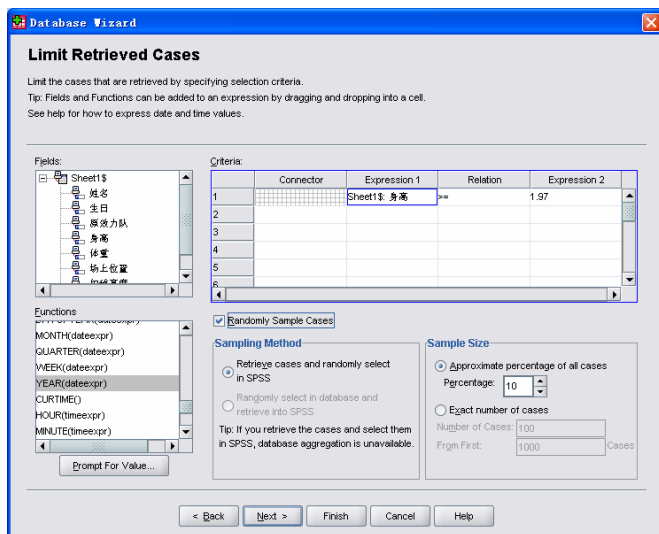


图 2-41 确定选择观测量的标准、方法和数量

• **Exact number of cases** 要求给出确切的采样数。可以在 **Number of cases** 后给出样品数量，由系统随机选择，也可以在 **From First** 后面给出数字，则系统从第一个观测值开始取给定数量的观测值组成样本。

单击 **Next** 按钮进入 **Define Variables** 对话框，见图 2-42。

(7) 在 **Defined Variables** 对话框中编辑变量属性。由于只能确定变量类型，一般数值型变量宽度采样用默认值。在最后一行确定字符串变量的长度。下设两个选项：

① **Width for variable-width string fields** 字符串变量的宽度在后面的数字栏内填写，否则按默认的 255 字节（255 个英文半角）确定宽度。

② **Minimize string widths based on observed values** 根据观测值的实际宽度确定字符串变量的最小宽度。

(8) 如果想把转换后的数据先保存起来，可以单击 **Next** 按钮，进入下一个对话框，否则可以直接单击 **Finish** 按钮，开始转换，转换结果见图 2-43。

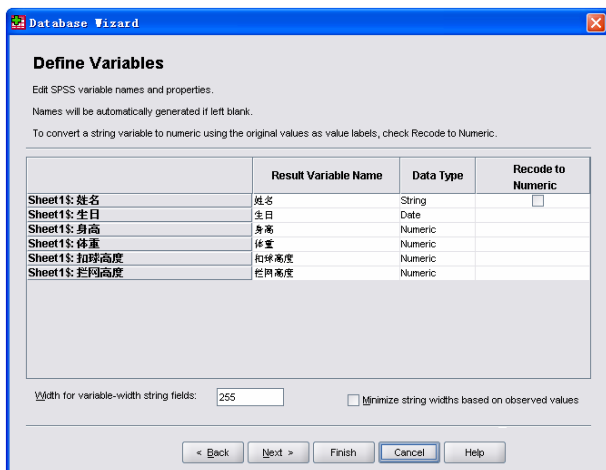
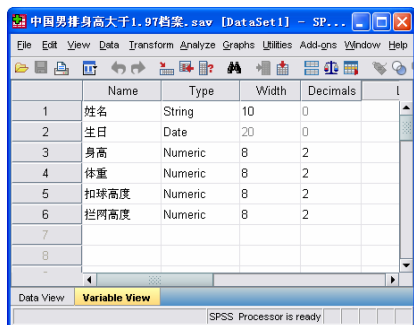


图 2-42 编辑变量属性对话框



(a)



(b)

图 2-43 中国男排数据库选择变量、观测值转换为 SPSS 数据文件的结果

**注意：**要重新编辑变量的测度类型。因为转换后的变各量测度类型不一定正确。

图 2-38~图 2-42 是将中国男排档案资料数据库（Excel 工作表）转换成 SPSS 数据文件的过程，只选择身高大于等于 1.97 的运动员的姓名、生日、身高、体重、扣球高度、



拦网高度 6 个变量进行转换。转换结果各变量均为 Norminal 标称变量，应该把身高、体重、扣球高度、拦网高度的测度类型改为 Scale 尺度型。

### 2.2.3 观测量的查重

1. 实际工作中有时会输入重复的数据

- 同一个观测量输入了多次；
- 多个观测量共用一个标识变量的值，但是第二标识变量的值不同；例如同一个家庭的多个成员，共用一个家庭地址或家庭编号；
- 标识变量值相同，非标识变量值不同；例如同一个人或同一个公司，多次或在不同时间购买不同的产品，在记录购买情况的数据文件中，这个人的名字或编号会出现多次。

2. 识别与处理重复观测量的方法

- (1) 按 Data→Identify Duplicate Cases 顺序单击菜单项，打开如图 2-44 所示对话框。
- (2) 定义识别重复观测量的根据。

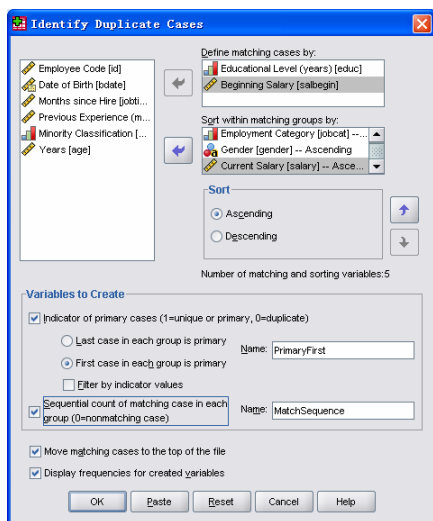


图 2-44 观测量查重主对话框

将源变量框中的识别变量移到右边的 Define matching cases by 框中，这些识别变量值相等的观测量被认为是重复的观测量。可以选择一个也可以选择多个。

系统将按第一个识别变量对数据文件排序，对第一个变量值相同的再按第二个识别变量排序。这两个变量的值都相同的观测量就是读者定义的重观测测量。

系统为标识重复观测量生成一个变量。对重复观测量，该变量的值为 0，非重复观测量其值为 1。

(3) 指定组内排序变量。

在源变量框内选择一个变量送入 Sort within matching groups by 框内。对符合同一重复条件的观测量组按该变量值排序。可以使用 Sort 右侧的上下箭头改变排序变量的顺序。

(4) 指定组内排序规则。

在 Sort 栏内的两个单选项中选择一个：Ascending 或 Descending，即按排序变量值升序还是降序排列组中的观测量。

如果指定了两个排序变量，则在 Sort 变量栏中选择一个，定义一次升序或降序。有几个排序变量就定义几次。系统默认按排序变量值的升序排列。

(5) 指定指针变量的特性。

对经过查重的数据文件，生成一个指针变量，在 Variables to Create 栏内指定该指针变量如何标识重复观测。

① Indicator of primary cases (1=unique or primary, 0=duplicate)，选择此项，产生的指针变量的值，对不重复的观测其值为 1，对重复的观测量中的主观测量，该变量值也为 1；对重复的观测量非主观测量（不满足 Primary 定义的），其值为 0。

② 定义重复观测量组中的主观测量的条件。

- Last case in each group is primary，选择此项后变量名为 Primary Last，在重复观测量组中，最后一个观测量的指针变量的值为 1，其他为 0。

- First case in each group is primary，选择此项后变量名为 Primary First 在重复观测量组中，第一个观测量的指针变量的值为 1，其他为 0。

- Filter by indicator values，选择此项后用指针变量作为过滤变量，非主重复观测量将从分析中去除，但无须从数据中文件删除。输出的结果和报告与这些观测量无关。也就是说，重复的观测量只留一个参与后续的分析，根据前两个选择项决定是保留排序后的重复观测量的第一个还是保留最后一个。

- Sequential count of matching cases in each group(0=nonmatching case)，选择此项，后面的 Name 栏显示产生另一个变量 MatchSequence，它对有  $n$  个重复观测量的组中各观测量标  $1 \sim n$  的值。每个重复观测量组自行排列。如果指定了排序变量，排列顺序取决于排序变量。如果没有指定排序变量，排列顺序取决于观测量在原始数据文件中的顺序。

- Move matching cases to the top of the file，选择此项，查重执行的结果会把有重复观测量的组移到数据文件的顶部，以便观察。

- Display frequencies for created variables，选择此项，要求生成频数表，包括所生成的新变量各值的计数。例如对主指针变量，频数表给出 0 值的个数和 1 值的个数。1 值的数目表明数据文件中共有多少个无重复的单一观测量和主观测量。

【例 5】数据文件 data02-01 为有 474 个观测量雇员情况的数据。变量有：id（雇员编号）、gender（性别）、bdate（出生日期）、educ（受教育年限）、jobcat（职务等级）、salary（当前工资）、salbegin（起始工资）、jobtime（雇用工作月数）、prevexp（以前的工作经历月数）、minority（民族）和 age（年龄）。

使用查重功能查看雇员受教育程度、职务的构成以及初始工资情况。操作步骤如下：

(1) 按 Data→Identify Duplicate Cases 顺序单击菜单项，打开对话框。

(2) 在源变量框中选择 educ、salbegin 作为识别变量移到 Define matching cases by 框中。

(3) 在源变量框中选择 jobcat、gender、salary 送入 Sort within matching groups by 框内，作为排序变量。

(4) 设置按 jobcat 降序，gender、salary 升序排列。

- (5) 定义重复观测量中开始的一个为主观测量，变量名为 PrimaryFirst。
- (6) 要求生成重复观测量的顺序变量，选择 Sequential count of matching cases 项，生成的新变量名为 MatchSequence。
- (7) 选择 Display frequencies for created variables 项，要求生成频数表。
- (8) 选择 Move matching cases to the top of the file 项，把有重复观测量的组移到数据文件顶部。

提交运行后，结果见表 2-2、表 2-3 和图 2-45。

表 2-2 显示，有重复的观测量共 312 个，主观测量 162 个；就是说，一共 162 组中有重复的测量组合。总观测量数是 474 个。

表 2-2 指针变量概况表

Indicator of each first matching case as Primary					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Duplicate Case	312	65.8	65.8	65.8
	Primary Case	162	34.2	34.2	100.0
	Total	474	100.0	100.0	

表 2-3 重复观测量的频数分布表

Sequential count of matching cases					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	94	19.8	19.8	19.8
	1	68	14.3	14.3	34.2
	2	68	14.3	14.3	48.5
	3	42	8.9	8.9	57.4
	4	27	5.7	5.7	63.1
	5	22	4.6	4.6	67.7
	6	21	4.4	4.4	72.2
	7	18	3.8	3.8	75.9
	8	16	3.4	3.4	79.3
	9	14	3.0	3.0	82.3
	10	13	2.7	2.7	85.0
	11	12	2.5	2.5	87.6
	12	10	2.1	2.1	89.7
	13	8	1.7	1.7	91.4
	14	7	1.5	1.5	92.8
	15	6	1.3	1.3	94.1
	16	5	1.1	1.1	95.1
	17	4	.8	.8	96.0
	18	3	.6	.6	96.6
	19	3	.6	.6	97.3
	20	3	.6	.6	97.9
	21	3	.6	.6	98.5
	22	3	.6	.6	99.2
	23	3	.6	.6	99.8
	24	1	.2	.2	100.0
Total		474	100.0	100.0	

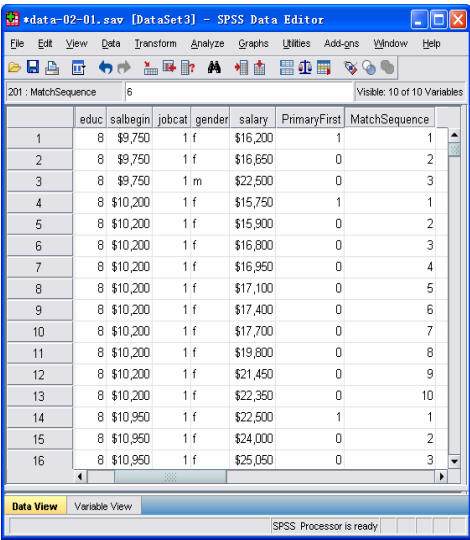


图 2-45 查重后数据排列

表 2-3 是频数分布表。举例说明各项内容：Valid 为 6 的 Frequency 是 21，Percent 为 4.4%，Cumulative percent 是 72.2%，其含义是重复的观测量数为 6 的组共有 21 组，占总观测量数的 4.4%，自不重复的观测量（Valid 为 0）到一组有 6 个重复观测量的观测

量总数占总数 474 的 72.2%。

图 2-45 是查重过程运行结束后的数据窗口。为便于查看，删去了与此例无关的变量。新变量 PrimaryFirst 和 MatchSequence 的含义及其各种值的含义也是显而易见的。

## 2.3 数据文件操作

### 2.3.1 数据文件的拆分与合并

#### 1. 数据文件的拆分

在进行数据处理时经常要对数据文件中的观测量进行分组分析，但有些分析功能没有设置对分组变量的选项。例如想使用 Descriptives 功能（Analyze descriptive 菜单中）分别求出男生、女生的平均身高。在进行分析之前必须对该数据文件进行拆分。这里的“拆分”并非将一个数据文件拆分为两个或若干个独立的数据文件，而是在同一个数据文件中按某个条件分组。若对数据文件进行了拆分处理，拆分处理一直有效，直到取消拆分处理或更改拆分变量后，才会有新的变化。关闭 SPSS，也会使拆分失效。具体操作步骤如下。

(1) 读取数据文件 data02-05。

(2) 按 Data→Split File 顺序打开 Split File 对话框，如图 2-46 所示。

(3) 根据对数据的具体需要选择以下选项。

- Analyze all cases, do not create groups, 对所有的数据进行处理，不产生分组。这是系统的默认选项。

- Compare groups, 将各分组的观测测量数据所得的结果放在一起进行比较。

- Organize output by groups, 按组输出。即分别显示各组所得的统计结果。

(4) 从左侧的源变量框中将一个或若干个要进行分组的变量名选入 Groups Based on 框中。此处最多可以选择 8 个变量作为拆分变量。这些变量所起的作用相当于排序的 By 变量。

如果只选择了一个变量，以后的分析将会依据该变量的每一个值分为一组，分别进行分析。例如选择性别变量 Sex，分析时分别按 Sex=0 和 Sex=1 把观测量分为两组进行分析。

如果选择了若干个变量，以后的分析将会依据所选择的变量各值的组合分组，对每个组分别进行分析。例如选择了变量 sex，它有两个水平：sex=0、sex=1；还选择了变量

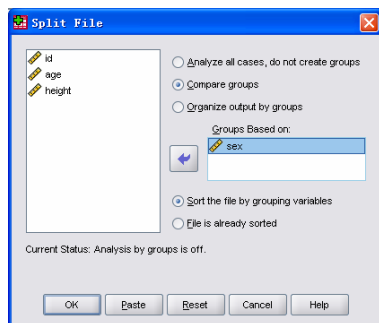


图 2-46 拆分数据文件对话框

age, 它有三个水平: age=11、age=12、age=13。分析时分为六组进行: sex=0, age=11; sex=0, age=12; sex=0, age=13; sex=1, age=11; sex=1, age=12; sex=1, age=13。

(5) 指明数据文件的当前状态。

① File is already sorted, 表示数据文件已经按所选择的变量排序。

② Sort the file by grouping variables, 表示要求按所选择的变量对数据文件进行排序,

作为拆分文件在分析时才起作用。

而从数据窗口中看上去, 经过拆分的数据文件与经过同样的变量排序的文件是相同的。如果在进行拆分之前进行了排序, 则会节省拆分所需的时间。

(6) 单击 OK 按钮执行并完成拆分。拆分结果见图 2-47。图(a)是按 sex 变量值拆分的结果。图(b)是按 sex、age 两个变量值拆分的结果。读者可以用 data02-05a 做实验。

(a) Data View window showing data split by 'sex'. The 'sex' variable has two categories: 0 and 1. The data is sorted by 'sex'.

(b) Data View window showing data split by 'sex' and 'age'. The 'sex' variable has two categories: 0 and 1. The 'age' variable has three categories: 11, 12, and 13. The data is sorted by 'sex' and 'age'.

(a)

(b)

图 2-47 选取不同拆分变量的拆分结果

## 2. 合并数据文件

### (1) 两种合并方式

合并数据文件是指将外部数据中的观测量或者变量合并到当前数据文件中去, 它包括两种合并方式。

① 从外部数据文件增加观测量到当前数据文件中。这种方法称为纵向合并或追加观测量。相互合并的数据文件中应该有相同的变量, 不同的观测量。

② 从外部数据文件增加变量到当前数据文件中, 称为横向合并。相互合并的数据文件中包含不同的变量。

### (2) 增加观测量 (Add Cases) 的纵向合并

① 首先在数据窗口中打开一个数据文件 data02-06, 如图 2-48(a)所示。与一个未打开的数据文件 data02-07 合并。Data02-07 的数据见图 2-48(b)。两个数据文件都有相同的变量 id、sex、age。

	id	sex	age	height
1	6	1	13	163
2	7	1	12	155
3	8	1	11	148
4	9	2	12	156
5	10	2	13	160

	id	sex	age	w	h
1	1	1	11	38	140
2	2	2	12	40	145
3	3	1	13	50	160
4	4	1	12	50	156
5	5	2	11	29	130

(a)

(b)

图 2-48 两个数据文件的原始状态

② 按 Data→Merge Files→Add Cases 顺序,

打开 Add Cases to data02-06.sav 对话框, 见图 2-49。指定一个要与之合并的数据文件。两种情况:

- An open dataset 框中列出与 data02-06 同时打开的数据文件, 可以从文件列表中选择一一个与之合并。

• An external SPSS data file 指定一个未打开的 SPSS 数据文件与 data02-06 合并。单击 **Browse** 按钮，指定一个外部 SPSS 数据集。

指定了与主文件合并的数据文件，单击 **Continue** 按钮，打开如图 2-50 所示的对话框。

③ **Variables in New Active Dataset** 框中列出的变量是在两个数据文件中变量名相同、类型相同的变量 (id、sex、age)。这些变量直接包括在合并后的新文件中。

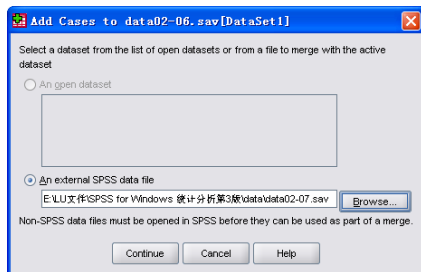


图 2-49 指定与主文件合并的数据文件

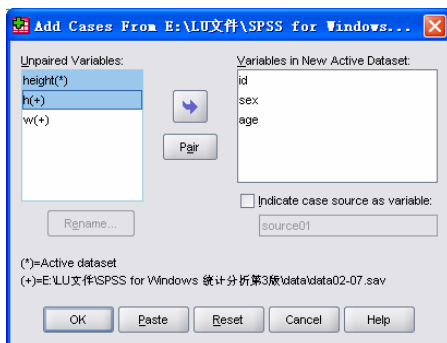


图 2-50 增加观测量对话框

**Unpaired Variables** 框中列出的变量是未配对变量，有 Height、h 和 w 这些在另一个数据文件中找不到变量名和类型与之相同的变量，即它们不能配对。标有“\*”的是当前数据文件中的变量，标有“+”的是外部数据文件中的变量。

#### ④ 根据情况处理数据

- 只合并两个数据文件中变量名和类型都相同的变量的观测量时，单击 **OK** 按钮。
- 追加外部数据文件中名称不同的变量（不匹配变量）的观测量。此时需要首先在 **Unpaired Variables** 框中设置配对变量，即先选取一个变量，再按住 **Ctrl** 键选取与之配对的变量，然后单击 **Pair** 按钮将它们送入新的数据文件变量表中，显示“height & h”最后单击 **OK** 按钮。没有配对的变量，也可以出现在合并有的数据集中，只要选择并送入右面的框中。

- 也可以改名以后再送入右框。在 **Unpaired variables** 中选择一个变量，单击 **Rename** 按钮，在 **Rename** 对话框中给出新名，单击 **Continue** 即可。

- **Indicate case source as variable** 指定一个变量，值为 1 表明来自工作数据文件的观测量，值为 0 表明是外部数据文件中的观测量。默认的变量名为 **Source01**，也可以自己命名。

**【例 6】**图 2-48(a)所示的当前工作数据文件中的变量 **height** 与外部文件数据文件中的变量 **h** 均为身高数据，如图 2-48(b)所示，只是变量名称不同。在未配对变量表中选择这两个变量，单击 **Pair** 按钮。在新工作数据文件的变量表中显示“height & h”。指定生

成指针变量, 使用默认名 Source01。单击 OK 按钮, 合并结果如图 2-51(a)所示。

未配对变量表中的变量在配对时要求一定具有相同的变量类型。宽度不相同, 当前文件中的变量宽度应当大于等于外部文件变量的宽度 (如 height 的宽度大于等于 h 的

宽度)。如果当前文件中的变量宽度小于外部文件变量的数据的宽度 (如 height 的宽度小于 h 的宽度), 在合并后外部文件被合并的观测量中的相应变量数据会丢失。若干个星号 “\*” 表示丢失的变量值。

对于只在一个数据文件中含有的变量 (例如变量 w 仅在外部的文件中存在), 如果不进行配对, 但要求包含在新的数据文件中, 只要选择这个变量, 并将其移入新

数据文件变量表中即可。图 2-51(b)是将 w 变量移入新数据文件变量表中, 但 w 变量并没有与之配对的变量, 由于当前工作数据文件不包括 w 变量, 因此相应的观测量 w 值为缺失值。

### 3. 增加变量 (Add Variables)

增加变量有两种方式:

- 两个数据文件按观测量顺序一对一地横向合并;
- 按关键变量合并, 即要求两个数据文件必须有一个共同的关键变量, 两个数据文件中关键变量值相同的观测量合并为一个观测量。

下面以 data02-08 为当前工作数据文件, 包括变量 id、sex、age、h、w, data02-09 为外部数据文件, 包括变量 id、w。以这两个数据文件横向合并为例, 说明操作步骤。

(1) 打开 data02-09 数据文件, 显示在另一个数据编辑窗口。

(2) 在 data02-08 数据编辑窗口, 按 Data→Merge Files→Add Variables 顺序, 打开 Add Variables to data02-08.sav 对话框。见图 2-52。在 An open dataset 栏内显示已经打开的数据文件的存储位置和文件名 data02-09。单击选择这个文件。然后单击 Continue 按钮。

(3) 在打开的如图 2-53 所示的 Add Variable from data02-09.sav 对话框中, 左栏 Excluded Variables 列出的是两个文件中的同名变量。只有这样的变量可以作为关键变量。对话框右侧的 New Active Dataset 矩形框中, 列出了可以在新工作数据文件中存在的变量。

Figure 2-51 consists of two side-by-side screenshots of the SPSS Variable View window. Both windows are titled '\*data02-06.sav [DataSet3] - S...'.  
 Screenshot (a) shows a table with 6 columns: id, sex, age, height, source01, and y1. The data rows are numbered 1 to 10. The 'y1' column contains values 0 or 1.  
 Screenshot (b) shows the same table after adding a new variable 'w'. The 'w' column contains values 38, 40, 50, 50, 29, and missing values (represented by dots) for the other rows. The 'source01' column remains the same as in (a).

(a)

(b)

图 2-51 不同变量情况的观测量合并结果



在两个矩形框中标有“\*”的是当前工作数据文件中的变量，标有“+”的是指定的外部数据文件或已经打开的另一个数据文件中的变量。

#### (4) 根据情况处理数据。

① 如果没有名字相同的变量，不用指定关键变量。要想合并两个数据文件中的变量，单击 OK 按钮即可开始横向合并两个数据文件了。结果是按观测量出现的顺序一对一地合并。

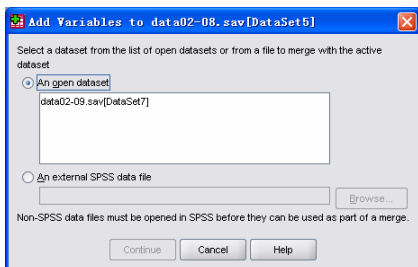


图 2-52 指定变量来源的数据文件

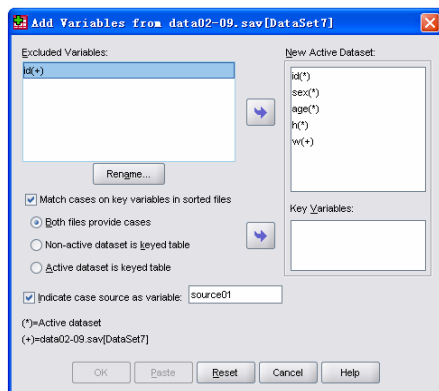


图 2-53 组织新数据文件中的变量

② 如果两个数据文件中具有同名的变量，那么合并的结果保留当前数据文件中同名的变量加上外部数据文件中不同名的变量。

③ 选择在当前数据文件与外部数据文件中包含的同名变量作为关键变量，需要先对数据文件按关键变量值的升序排序。

对于排序后关键变量值相同的合并为一个观测量。图 2-54(a)为当前数据文件 data02-08，图 2-54(b)为第 2 个数据文件 data02-09，均已经按关键变量 id 排序。图 2-54(c)为合并后的数据文件。观测量 id=60, 65, 68 的观测量，在两个数据文件中都存在，横向合并。

对于两个文件中关键变量值不同的观测量处理方法是，选择 Match cases on key variables in sorted files，激活下面三个选项。在以下 3 个选项中选择一种处理方式。

Both files provide cases，即观测量由两个数据文件提供。合并的结果是将第 2 个数据文件的观测量追加到当前工作数据文件中，如图 2-54(c)中的 id=60, 64, 65, 67、68 的观测量。与 data02-08 的 id 值相同的 id=60、65、68 合并，id 值不相同的 id=64、67，也追加到 data02-08 文件中，结果保存在 data02-08a.sav 中。

Non-active dataset is keyed table，即保持当前数据文件中的观测量数目不变。在第 2 个数据文件中，只有那些与当前数据文件中关键变量等值的观测量才能合并到工作数据文件中，例如 data02-08 与 data02-09 数据文件以这种方式合并，其结果见图 2-55(a)。



Active dataset is keyed table, 当前数据文件中的观测量按与第 2 个文件中的关键变量值相等时并入第 2 个文件，例如 data02-08 与 data02-09 数据文件以这种方式合并，其结果见图 2-55(b)。

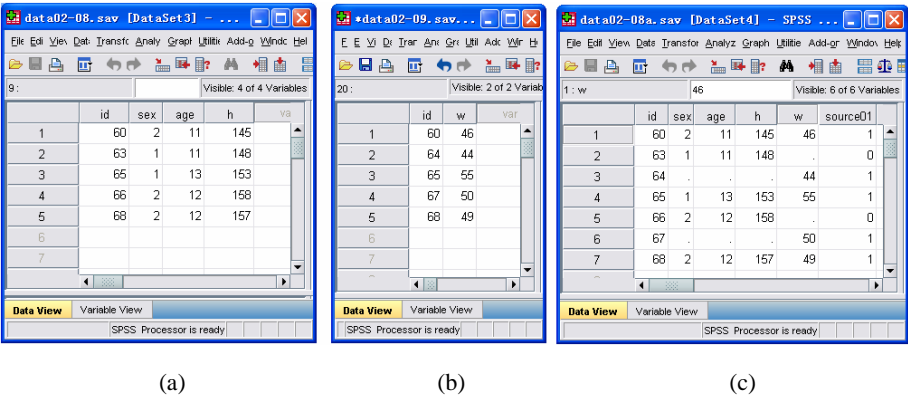


图 2-54 由两个排序数据文件提供合并数据

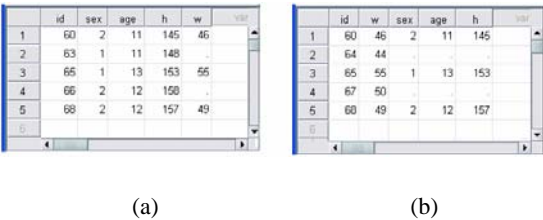


图 2-55 以关键变量值相等的原则合并

最后将在 Excluded Variables 框中选择的关键变量 (id)，通过单击下面一个向右箭头按钮移到 Key Variables 框中。单击 OK 按钮将指定条件和方式的合并提交系统执行。系统将提示警告：如果两个文件没有按关键变量排序，合并可能失败。因此，在执行合并功能之前，必须将内、外两个文件均按关键变量排序。

(5) 几点说明。

- ① 如果在当前数据文件中与外部数据文件中有同名的变量，外部数据文件中的变量列于 Excluded Variables 框中，当前数据文件中的变量列于右面的 New Active Dataset 框中。
- ② Excluded Variables 框中的变量若选为关键变量，则可以将其移到 Key Variables 框中。与其同名的 New Active Dataset 框中的变量消失。
- ③ 如果一定要将 Excluded Variables 框中外部数据的同名变量合并到新的数据文件中，那么应先为该变量更名；即单击 Rename 按钮，在被打开的相应对话框中赋予该

变量一个新名。然后选择该变量，并单击上面一个向右箭头按钮将其移到 **New Active Dataset** 框中。

④ **New Active Dataset** 框中的变量均为新数据文件中的变量。如果不想使某变量出现在框中，则选择这个变量，将其移到 **Excluded Variables** 框中。

⑤ 为变量更名。如果两个数据文件中有同名变量，但内容不同，需要对其中一个变量更名；如果两个文件中作为关键变量的两个变量不同名，应该改成相同的变量名。

⑥ 生成新变量。如果选中 **Indicate case source as variable**，即显示数据来源变量，一个新的变量（读者输入的变量名称）将会加入到当前数据文件中。其变量值 0 表示观测量来自当前数据文件，1 表示观测量来自非工作数据文件。

## 2.3.2 观测量的排序与排秩

### 1. 观测量排序

在进行数据处理过程中，有时需要按照某个或某些变量（排序变量）的值的顺序重新排列观测量在数据文件中出现的先后顺序，可以按下述步骤实现。

(1) 按 **Data→Sort Cases** 顺序打开 **Sort Cases** 观测量排序对话框，如图 2-56 所示。

(2) 在左侧的源变量框中选择排序变量，移到右侧的 **Sort by** 框中。

如果选择了两个以上的排序变量，列于首位称为第一排序变量，其后的顺序分别称为第二排序变量、第三排序变量……排序的结果与排序变量在 **Sort by** 框中的顺序有关。

排序的结果是观测量先按第一排序变量的值排列观测量，在第一排序变量的值相等的观测量组，按第二观测量的值排序，以此类推。

如果排序变量是字符型的，英文排序按拼写的字母 **ASCII** 码顺序排列。中文排序按拼音字母的 **ASCII** 码顺序排列。

(3) 确定排序的方式，即根据变量顺序进行排列。

① 在 **Sort by** 框内选择一个排序变量。

② 在 **Sort Order** 栏内选择以下一种排序方式：

- **Ascending**，按所选择的排序变量的升序排列；
- **Descending**，按所选择的排序变量的降序排列。

(4) 重复第 3 步的操作可以指定下一个排序变量的排序方式。

(5) 单击 **OK** 按钮，即可完成排序工作。按 **Paste** 键可以在 **Syntax** 语句窗口中生成程序语句。

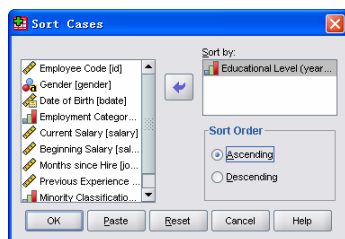


图 2-56 观测量排序对话框

## 2. 根据变量的值对观测量排序

在当前数据文件中产生秩变量的操作步骤如下:

(1) 按 Transform→Rank Cases 顺序单击菜单项, 打开 Rank Cases 对话框, 见图 2-57。

(2) 在左侧的源变量框中选择至少一个变量进入右侧的 Variable(s)框中。对每个变量产生一个秩变量。

(3) 在 Assign Rank 1 to 栏中选择秩的排列方式:

- Smallest value, 定义 1 为最小的数值的秩;
- Largest value, 定义 1 为最大的数值的秩。

(4) 读者可以选择一个或多个分组变量进入 By 框中, 系统将按 By 变量值分组排序。

(5) 单击 Rank Types 按钮打开如图 2-58 对话框, 指定产生的秩的算法:

① Rank, 简单秩。数据文件中的新变量值就是秩, 这是默认方式。新变量名为据以排秩的变量名前冠以“r”。

② Savage score, 原始分数秩。秩变量的值是依据指数分布所得原始分数。新变量名为原变量名前冠以“s”。

③ Fractional rank, 分数秩。新变量的值等于简单秩除以非缺失观测量的加权之和。

④ Fractional rank as percent, 百分比小数秩。秩值为其秩除以所有合法值的观测量数目之和乘以 100。

⑤ Sum of case weights, 观测量加权之和。新变量的值是观测量权重之和。在同组中新变量值是个常数。

⑥ Ntiles, 分段排序。在参数框中输入分段数, 分段数必须是大于 1 的整数。某一观测量的秩值是按该观测量占的百分位数的位置来决定的。例如: 如果输入的数值为 4, 那么变量值的百分位数低于 25% 的观测量的秩将被赋值为 1, 位于 25%~50% 的观测量的秩将被赋值为 2, 位于 50%~75% 的观测量将被赋值为 3, 高于 75% 观测量被赋值为 4。

⑦ Proportion estimates, 比例估计选项, 是与一个特别秩的分布的累计比估计。

⑧ Normal scores, 正态分数选项, 即与估计累计比相应的 Z 分数。

选择了⑦、⑧比例估计类型后, 激活下面的选项, 可以进一步指定计算公式, 即在 Proportion Estimate Formula 栏中进行选择。

• Blom, 由公式  $(r-3/8)/(w+1/4)$  决定, 其中  $r$  为秩,  $w$  为观测量的权重之和。此项为默认设置。

• Tukey, 由公式  $(r-1/3)/(w+1/3)$  决定, 此处  $w$  为观测量权重之和,  $r$  为秩。

• Rankit, 由公式  $(r-1/2)/w$  决定,  $w$  为观测量权重之和,  $r$  为序列, 范围  $1 \sim w$ 。

• Van der Waerden 选项, 由公式  $r/(w+1)$  决定, 此处  $w$  为观测量权重之和,  $r$  为秩。要求这是默认设置, 可单击此项, 输出窗口不显示这些信息。

(6) 确定结的秩。

变量值相同的, 称为结。结的秩次的决定原则可以在 Rank Cases: Ties 对话框中指定。

在主对话框中单击 **Ties** 按钮，打开结对对话框，如图 2-59 所示。

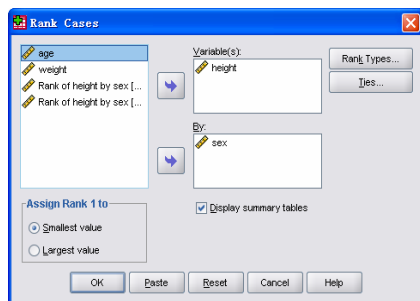


图 2-57 观测量排秩对话框

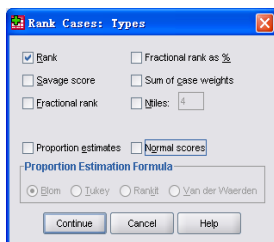


图 2-58 秩类型对话框

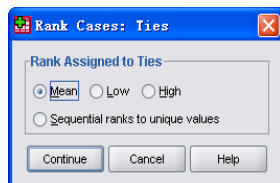


图 2-59 结点秩的确定

- ① **Mean**，相同值的秩取平均值。
- ② **Low**，相同值的秩取最小值。
- ③ **High**，相同值的秩取最大值。
- ④ **Sequential ranks to unique values**，相同值的秩取第一个出现的秩次值，其他观测值秩次顺序排列。体育比赛常用这种方法排列名次。

上述四种不同方法排秩的比较见表 2-4。

表 2-4 四种不同方法排秩的比较

观测量值	Mean	Low	High	Sequential
1.00	1	1	1	1
2.50	3	2	4	2
2.50	3	2	4	2
2.50	3	2	4	2
3.50	5	5	5	3
3.75	6	6	6	4

(7) 以上选项确定后，单击 **OK** 按钮，根据指定的变量、分组变量及其他选项计算秩，并生成新变量。在输出窗口中显示新变量的名称、标签、秩类型等总结性的信息。

### 2.3.3 对变量值重新编码

把连续变量变成分类变量或重新分类时需要重新编码。Transform 菜单中的 **Recode** 命令和 **Automatic Recode** 可以对多个类型相同的变量重新编码，生成新变量。新变量的值是重新编码的结果，也可以用新代码代替原始变量。**Automatic Recode** 命令是自动重新编码；**Recode** 命令允许在编码过程中进行人为干预。

#### 1. 使用 Recode 命令重新编码

Transform 菜单中有两项与重新编码有关：

- **Recode into Same Variables**，对一个变量重新编码，结果代替该变量；

• **Recode into Different Variables**, 生成新变量, 变量的值是编码的结果, 对话框如图 2-60 所示。

两个选项的主对话框区别仅在于 **Recode into Same Variables** 对话框中没有定义输出变量的部分。因此, 在此只叙述 **Recode into Different Variables** 生成新变量的操作。以 data02-10 为例说明对年龄 age 变量重新编码方法。

(1) 从变量列表中选择要重新编码的变量, 送入 **Numeric Variable→Output Variable** 框中。

(2) 每选择一个变量, 就在 **Output Variable** 的 **Name** 栏内输入新变量名, 在 **Label** 栏内输入新变量标签。单击 **Change** 按钮。

(3) 可以单击 **If** 按钮展开相应对话框, 根据条件选择要编码的观测量。

单击 **Old and New Values** 按钮, 展开 **Recode into Different Variables: Old and New Values** 对话框, 见图 2-61。左面 **Old Value** 是给出原变量值或值范围的区域, 每选择一项就在右边的 **New Value** 栏中选择一项, 或选择一项同时给出新变量的值。单击 **Add** 按钮, 将新、老变量之间关系, 即变量值与编码的对应关系送入 **Old→New** 栏中。

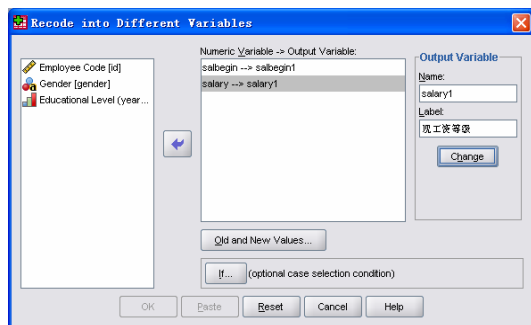


图 2-60 重新编码到新变量对话框

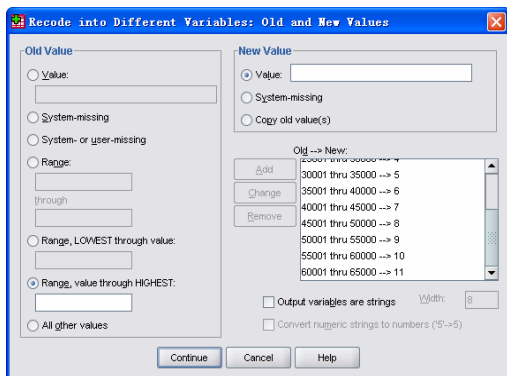


图 2-61 定义新变量值对话框

① **Old Value** 栏中选择并给出原始变量的值或值范围。

- **Value** 栏中输入单个值。
- **System-missing**, 为系统缺失值。
- **System- or user-missing**, 为系统缺失值或用户缺失值。
- **Range: \_\_through\_\_** 框, 在两个输入区中给出最低和最高两个值, 定义这个区间内的所有值。
- **Range, LOWEST through value** 框, 给出一个值, 定义小于等于这个值范围内的值。
- **Range, value through HIGHEST** 框, 给出一个值, 定义大于等于这个值范围内的值。
- **All other values**, 定义前面所有定义没有包括的值。

② New Value 栏中, 针对 Old Value 栏中给出的值, 给出新代码。

- 在 Value 框针对 Old Value 给出的值, 输入对应的新代码的值。
- System-missing, 将 Old Value 给出的值定义为缺失值。
- Copy old value(s), 新代码与 Old Value 给出的值相同。

**注意:** 各段值的衔接, 表达式一定不要漏掉某些介于两组值之间的值。

如果要按不同情况分组定义, 还应该单击 if 按钮进入对话框, 阐明条件。有关操作见本章根据已经存在的变量生成新变量一节。

(4) 定义结束, 单击 Continue 按钮, 返回主对话框, 单击 OK 按钮。

**【例 7】重新编码实例。**

打开数据文件 data02-11。

(1) 编码要求: 对职工的起始工资 salbegin 和当前工资 salary 重新编码, 对应的新变量名分别为 salbegin1 和 salary1 将连续变量编码为分类变量。要求的代码见表 2-5。

(2) 操作要点

① 按 Transform→Recode into Different Variables 单击菜单项, 打开对话框, 将 salary 和 salbegin 送入 Numeric Variable→Output Variable 框中。

单击 salary 在 Output variable 栏的 Name 后输入新变量名 salary1 和在 Label 后输入标签“当前工资等级”, 单击 Change 按钮, 则在 Numeric Variable→Output Variable 框中显示 salary→salary1。

表 2-5 编码表

salary 和 salbegin	<=16000	16001~ 20000	20001~ 2500	25001~ 3000	30001~ 35000	35001~ 40000	40001~ 45000	45001~ 50000	50001~ 55000	55001~ 60000	60001~ 65000	>=65001	System- missing
salary1 和 salbegin1	1	2	3	4	5	6	7	8	9	10	11	12	System- missing

再单击 salbegin, 在 Output variable 栏的 Name 后输入新变量名 salbegin1 和在 Label 后输入标签“起始工资等级”, 单击 Change 按钮。Numeric Variable→Output Variable 框中显示 salbegin→salbegin1。

单击 Old and New Values 按钮, 展开相应对话框, 定义新旧变量对应关系。

② Old value 栏中选择 Range LOWEST through Value, 输入 16000, 在 New value 栏的 value 中输入 1。

③ Old value 栏中选择 Range \_\_through\_\_, 输入值 16001 和 20000, 在 New value 栏的 value 中输入 2。新变量代码为 3~11 的都与此操作相同。

④ Old value 栏中选择 Range, Value through HIGHEST, 输入 65001, 在 New value 栏的 value 中输入 12。

⑤ Old value 栏中选择 System-missing, 在 New value 栏的 value 中也选择 System-missing。

定义完成, 单击 Continue 按钮, 返回主对话框, 单击 OK 按钮, 提交运行, 结果见

数据盘中的 data02-11a.sav。

(3) 对于等级较多的重新编码, 写个小程序会更简单。本例运行程序如下:

```
RECODE salary salbegin  
  (SYSMIS=SYSMIS) (Lowest thru 16000=1) (16101 thru 20000=2)  
  (20001 thru 25000=3) (25001 thru 30000=4) (30001 thru 35000=5)  
  (35001 thru 40000=6) (40001 thru 45000=7) (45001 thru 50000=8)  
  (50001 thru 55000=9) (55001 thru 60000=10) (60001 thru 65000=11)  
  (65001 thru Highest=12) INTO salary1 salbegin1 .  
VARIABLE LABELS salary1 '工资等级' /salbegin1 '初始工资等级'.  
EXECUTE .
```

程序由两部分组成。

第一段程序是 RECODE 过程语句。RECODE 是命令关键字, 后面是要进行重新编码的原始变量。中间部分是编码规则的表达式。最后 INTO 跟着两个新变量名。

每个编码表达式由等号连接两部分, 等号前是原始变量值或值范围表达式, 等号后面是新代码值。表达式有几种形式: (SYSMIS=SYSMIS) 表示新变量的系统缺失值与原始变量定义相同; (Lowest thru C1=C3) 定义原始变量值小于等于 C1 的, 新变量的值为 C3; (C1 thru C2=C3) 定义原始变量值在 C1 与 C2 之间的, 包括 C1、C2, 新变量的值为 C3; (C1 thru Highest=C3) 定义原始变量的值大于等于 C1 的, 新变量的值为 C3。

第二段程序是 VARIABLE LABELS 过程语句, 为新变量加变量标签。VARIABLE LABELS 是过程语句关键字, 后边是变量名与变量标签, 中间用空格分隔。

## 2. 使用 Automatic recode 自动重新编码

(1) 使用自动编码功能对数据进行预处理的需要。

① 原始分类变量的分类值不是等间隔的, 会在进行频数分布分析时形成空单元, 不但浪费计算机资源, 也使输出表格臃肿, 不利于得出结论。

② 有些分析过程要求参与分析的分类变量必须是数值型的, 不能是字符型的。需要转换。某些分析过程要求分类变量值是整数。

(2) 以 data02-11 中的受教育程度变量 educ 为例, 说明自动重新编码的操作。

① 按 Transform→Automatic Recode 顺序打开相应的对话框, 见图 2-62(a)。

将要自动编码的变量 educ 送入右面的 Variable→New Name 栏, 显示 educ→???????, 在下面的 New Name 栏输入新变量名 educ1, 单击 Add New Name 按钮, Variable→New Name 栏中显示新旧变量名对应关系: educ→educ1。

② 在 Recode starting from 栏中选择 Lowest value 表示从最小值开始编码。也可以选择从最大值开始编码, 但是对受教育程度这个有序分类变量来说最好新编码顺序与原来的受教育年限的原始值一致。单击 OK 按钮。在输出窗口显示编码结果, 如图 2-62(b) 所示。

③ 输出结果表明, 原始值从 8~21, 缺少 9、10、11、13, 新变量值从 1~10, 使用原始值作为值标签。如果在分析输出表中使用值标签, 就可以得到比较满意的、易于解释的结果。

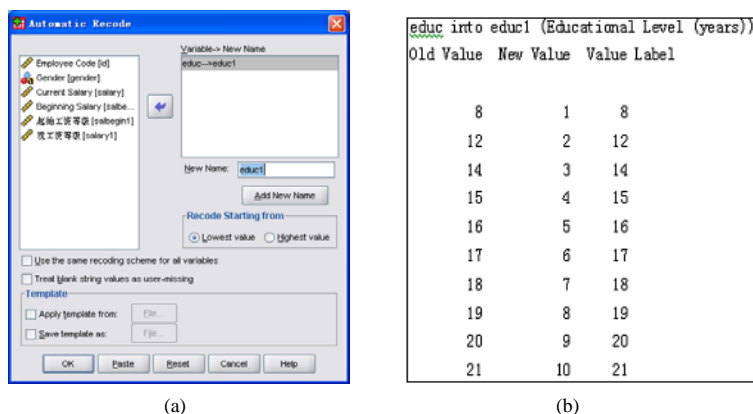


图 2-62 自动编码对话框和输出结果

(3) 对话框中还有几个选项:

① Use the same recoding scheme for all variables, 对所有变量使用这同一个重新编码方案。

② Treat blank string valuse as user-missing, 把空字符串变量值作为用户缺失值处理。

③ Template 模板选项。

- Apply template from: 选择此项, 指定一个模板文件作为设置本数据文件指定变量重新编码的模板。

- Save template as: 把当前的重新编码方案作为模板保存起来, 以便以后用在其他变量的重新自动编码上。

## 2.3.4 数据文件的转置与重新构建

分析工作中要求的数据排列方式往往与当前数据文件中的数据排列方式不同。为了满足分析过程对数据文件结构的要求, 就需要进行变换。使用移动、复制固然可以达到目的, 但是往往容易出错。本节介绍由变换工具自动变换的方法。

### 1. 数据文件的转置

利用数据的转置功能, 可以将数据文件中原来的行变成列, 原来的列变成行; 将观测测量转变为变量, 将变量转变为观测测量; 在新文件中建立一个其值为原来变量名的变量。转置后的数据文件与原来的数据文件完全不同, 应该保存到另一个文件名下。

操作步骤如下:

① 按 Data→Transpose 顺序打开 Transpose 对话框, 如图 2-63 所示。



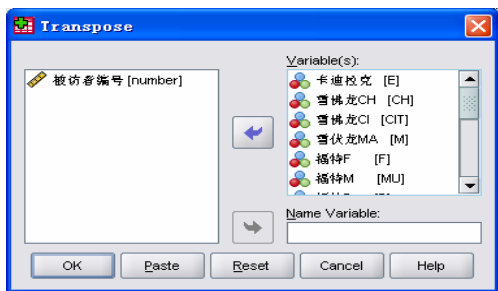


图 2-63 数据转置对话框

② 在左侧的源变量框中选择要进行转置的变量，用向右箭头按钮送到 Variable(s)框中。

这些变量在新数据文件中变成观测量（从列变成行）。新数据文件不会出现未被选择的变量。

③ 从源变量框中选择一个变量送入 Name Variable 框中。该变量的值在新数据文件中作为变量名出现。一般选择标识观测量的变量，如观测量的编号、姓名等。

如果它是一个数值变量，新变量名为该变量各值冠以字母“K\_”。如果不选择 Name Variable，系统会自动给转置后的变量赋予 var001、var002、…、var00n 的变量名称。

④ 单击 OK 按钮，进行转置。Paste 按钮可以将相关语句粘贴到 Syntax 语句窗口中。

【例 8】数据文件转置操作实例。

data02-12 数据文件是对汽车市场调查的数据。25 名被访者对凯迪拉克、雪铁龙等 17 个品牌的汽车打分的结果数据，18 个变量为 17 个汽车名称和一个被访者编号。每个观测量就是一个被访者给 17 种车的打分。为了进行顾客偏好分析，需要对数据文件进行转置。转置结果经整理保存到 data02-12a 中。以此数据为例，说明转置操作。

本例选择 17 种品牌变量作为要转置的变量送入 Variable(s)框中。本例选择被访者编号变量 number 送入 Name Variable 框中。

在主对话框中单击 Paste 按钮，在语句窗口中生成如下程序：

FLIP

VARIABLES=E CH CIT M F MU P A CI CO G H V FI D R DL

/NEWNAME=number.

程序调用 FLIP 过程，VARIABLES 语句在关键字后面用等号连接变量表，各变量名用空格隔开。这些变量转置后变为观测量。NEWNAME 子命令指定一个标识变量。

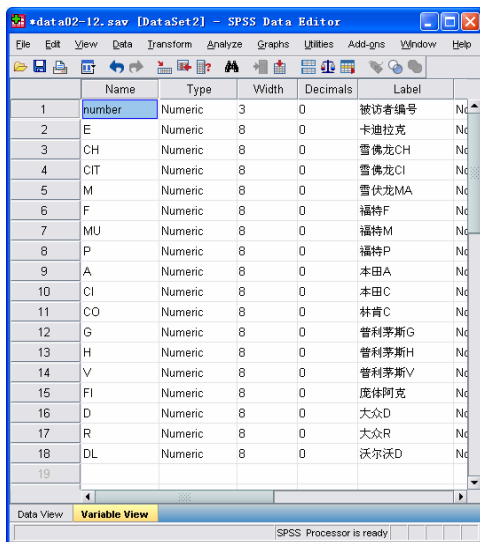
图 2-64(a)是转置前的变量观察窗口，变量是编号和 17 个品牌的汽车；图 2-64(b)为转置后的变量观察窗口，变量为 k\_1~k\_25 和系统自动生成的变量 CASE\_LBL。

在输出窗口中列出了所有新变量名。另外，如果数据包含缺失值，那么 SPSS 将其设置为系统的缺失值。为了保留缺失值，可以重新改变对变量中缺失值的设置。图 2-65 为转换后的数据观察窗，原来的 17 个变量转换成 17 个观测量。

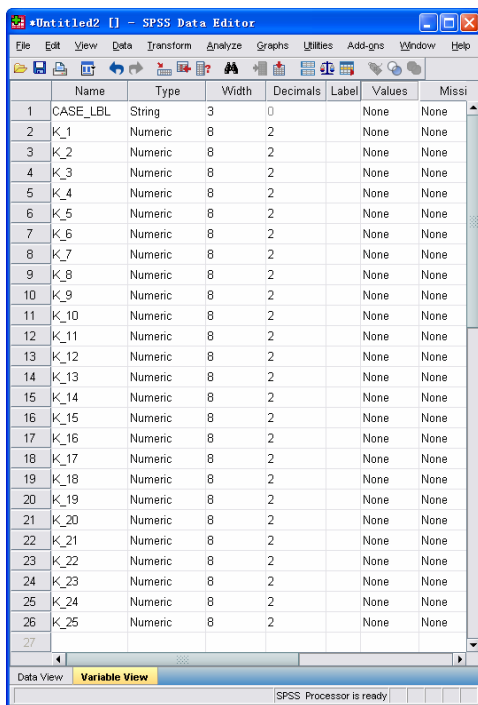
## 2. 数据文件的重新构建

### (1) SPSS 数据文件结构

SPSS 数据分析所需要的数据文件在数据观察窗中结构分为三种。



(a)



(b)

图 2-64 转置前后的变量观察窗口

	CASE_LBL	K_1	K_2	K_3	K_4	K_5	K_6	K_7	K_8	K_9	K_10	K_11	K_12	K_13	K_14	K_15	K_16	K_17	K_18	K_19	K_20	K_21	K_22	K_23	K_24	K_25
1	E	8	0	0	7	9	9	0	4	9	1	2	4	0	5	0	8	9	7	1	0	9	3	8	0	9
2	CH	0	0	5	1	2	0	0	4	2	3	4	5	1	0	4	3	0	0	3	5	1	5	6	9	8
3	CIT	4	0	5	3	3	0	5	8	1	4	1	6	1	6	4	3	5	4	4	7	4	7	7	9	5
4	M	6	0	2	7	4	0	0	7	2	3	1	2	1	3	4	5	5	4	5	6	6	8	6	5	8
5	F	2	0	2	4	0	0	6	7	1	5	0	2	1	4	4	3	5	3	0	6	4	8	6	5	5
6	MU	5	0	0	7	1	9	7	0	5	0	2	1	1	0	1	8	5	0	6	5	7	5	5	5	5
7	P	0	0	2	1	2	2	2	3	0	3	0	3	0	2	0	1	5	0	0	5	1	4	0	7	8
8	A	5	9	5	6	8	9	7	6	0	9	6	9	9	9	5	2	9	9	8	9	7	5	0	7	8
9	CI	4	8	3	6	7	0	9	5	0	7	4	8	8	8	5	2	5	6	7	7	6	5	0	7	5
10	CO	7	0	0	8	9	9	0	5	9	2	2	3	0	4	0	9	9	6	2	0	9	1	9	0	9
11	G	7	0	0	6	0	0	0	4	3	4	1	0	1	1	0	7	3	3	3	4	5	8	7	0	8
12	H	3	0	0	5	0	0	5	6	3	5	4	6	1	3	0	2	4	4	4	6	7	5	6	5	5
13	V	4	0	0	5	0	0	3	6	1	4	0	2	1	6	0	2	7	5	4	4	7	6	5	5	5
14	FI	0	1	0	7	8	9	5	6	1	3	2	0	1	2	0	6	9	5	8	2	6	5	9	0	7
15	D	4	8	5	8	6	9	6	5	0	8	8	7	7	7	9	5	3	7	7	8	9	5	0	0	0
16	R	4	8	5	8	5	0	9	7	0	9	6	9	5	7	9	5	4	8	7	8	8	5	0	0	0
17	DL	9	9	8	9	9	9	8	9	0	9	9	9	9	9	8	7	9	8	9	9	1	9	0	0	0

图 2-65 转置后的数据观察窗

① 简单数据文件。一个变量占一列，一个观测量占一行。例如对一个班的所有学生进行一项测试，所有分数仅出现在一列中，每个学生占一行。

② 有关一个观测量的信息占不止一行。例如，一个因素的每个水平占一行或不止一行，见表 2-6。一个因素的若干水平称作为一个观测量组，SPSS 数据分析中，当数据用这种方式构造时，因素经常是作为分组变量。

③ 有关一个变量的信息占不止一列。例如，一个因素的每个水平占一列，见表 2-7。一个因素的若干列称作为一个变量组。在 SPSS 数据分析中，当数据按这种方式构造时，因素常常涉及重复测量。

表 2-6 观测量组结构

factor	var
1	3
1	8
1	6
2	5
2	9
2	4

表 2-7 变量组结构

var1	var2
4	6
8	5
7	9

(2) 各种分析方法所需要的数据文件结构

① 要求观测量组数据结构的分析过程。数据必须按观测量组构建，以便做分组变量的分析。例如 General Linear Model 中的 Univariate 单因变量方差分析、Multivariate 多因变量方差分析、Variance Components 方差成分分析；Mixed Models 混合模型；OLAP Cubes 和独立样本 T 检验或 Nonparametric Tests 非参检验。

② 要求变量组数据结构的分析过程。数据必须按变量组构建，以便分析重复测量。例如 General Linear Model 的 Repeated Measures 重复测量，Cox 回归分析中的时间为因变量的协方差分析、配对样本 T 检验或相关样本的非参检验。

如果选择的分析过程所需要分析的数据结构与当前数据文件中的结构不符，需要进行变换。这个工作可以由 Data 菜单中的 Restructure 功能来完成。

(3) 变量组结构到观测量的转换步骤

- ① 单击 Data→Restructure...打开 Restructure Data Wizard 对话框，如图 2-66 所示。这是一个向导式的操作，主对话框中有三个选项，对应三种重新构建的类型。
- Restructure selected variables into cases, 将选择的变量转换成观测量。如果当前数据文件的结构是变量组的，要转换成观测量组结构，选择此项。
  - Restructure selected cases into variables, 将选择的观测量转换成变量。如果当前数据文件的结构是观测量组的，要转换成变量组结构，选择此项。
  - Transpose all data, 对所有数据进行转置。如果选择此项，关闭如图 2-66 所示对话框，打开如图 2-63 的 Transpose 对话框，进行转置操作。

② 一个变量组的结构转换成观测量组结构。以数据 data02-13-1 为例是 5 个学生，学号分别为 1~5；A、B、C 三门课程的考试分数变量：scoreA、scoreB、scoreC 和他们的身高 h、体重 w 和学号 no 的记录，如图 2-67(a)所示。在向导主对话框中选择第一项，单击 Next 按钮，打开如图 2-67(b)所示对话框。在对话框中选择 One，即要把一个变量组 scoreA、scoreB、scoreC 转换成一个观测量组。如果有两组变量要同时进行转换，则

应该选择第二项，More than one。

③ 单击 Next 按钮，打开如图 2-68 所示的对话框。

- Case Group Identification 栏，要求确定在新数据文件中的标识变量，其下拉列表中有三个选项：

- Use case number，使用观测

量顺序号；

- Use selected variable，使用选择的变量。

- None，不用标识变量。

本例有学号作为观测量标识，所以选择 Use selected variable，在 Variables in the Current File 栏中选择变量学号 no 送入右面的 Variable 栏中。

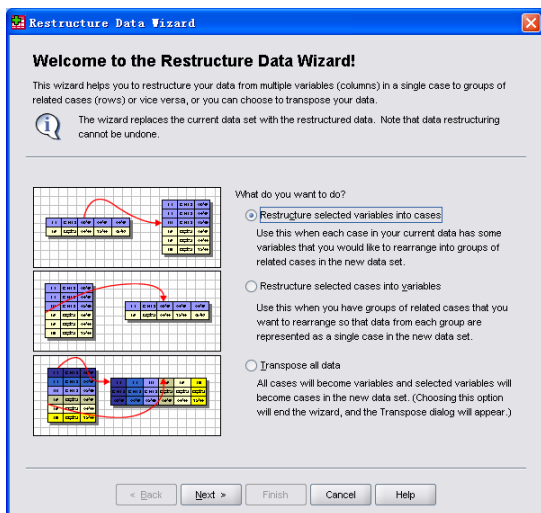
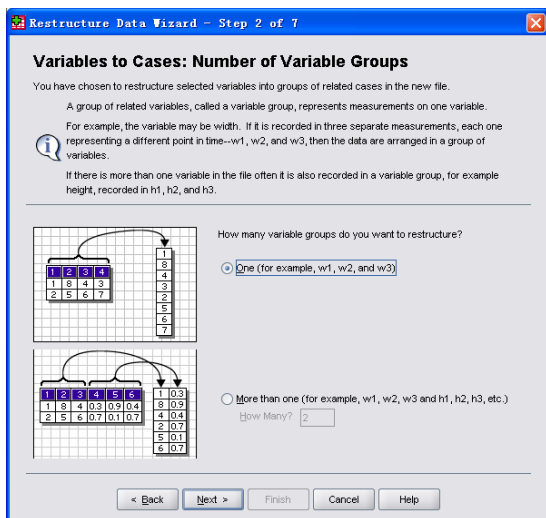


图 2-66 重建数据结构向导主对话框

	no	scoreA	scoreB	scoreC	h	w	var
1	1	99.5	93.5	81.5	169	60	
2	2	63.5	74.5	79.5	173	72	
3	3	69.0	86.5	81.0	174	65	
4	4	85.0	97.0	81.0	181	74	
5	5	78.0	67.5	69.5	167	60	
6							

(a)



(b)

图 2-67 原始数据与第 2 步对话框

- Variable to be Transposed 栏，确定要转换的变量。在当前数据文件中的变量栏内选择要转换的变量送入 Target Variable 下面的栏中。本例选择 scoreA、scoreB、scoreC。在 Target Variable 右面输入新文件中的变量名 score。

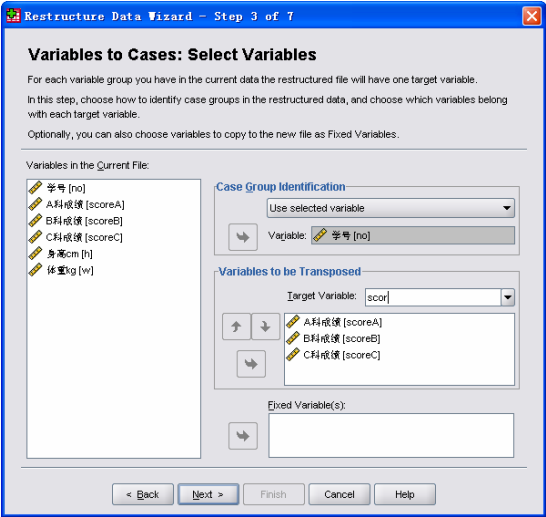


图 2-68 变量组转换到观测量组第 3 步对话框

中，索引变量可以作为分组变量，其下有三个选项：

- One，在大多数情况下，一个索引变量就足够了；本例选择此项。
- More than one。如果在当前文件中的变量组表现了多个因素的水平，可能要多个索引变量。输入想要产生索引变量的数目。

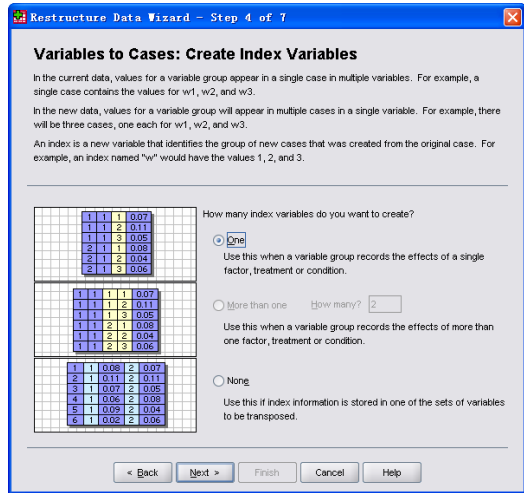


图 2-69 变量组转换到观测量组第 4 步对话框

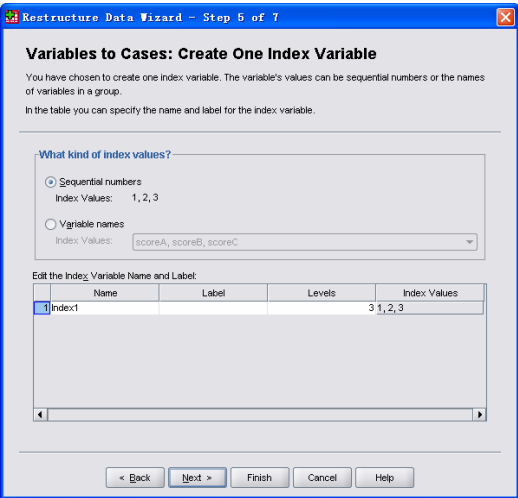


图 2-70 变量组转换到观测量组第 5 步对话框

- None，如果不需要生成索引变量指定这一项。

所指定的数目会对 Next 有影响，在下一步 Wizard 自动生成指定数目的索引变量。

• 在当前数据文件中变量栏中选择不进行转换，但还要出现在新文件中的变量，送入 Fixed Variable(s)栏中。

本例还有身高 h、体重 w 两个变量在分析中不会用到，不希望出现在转换后的文件中，所以这项没有操作。

④ 单击 Next 按钮，打开对话框如图 2-69 所示。这一步，决定是否在新文件中生成索引变量。索引变量根据原始变量组，在新文件中给它们按顺序编码。

How many index variables do you want to create 询问在新文件中应该生成多少个索引变量。在 SPSS 过程

⑤ 单击“Next”按钮，打开如图 2-70 所示的对话框。在这一步，确定索引变量的值。在 What kind of index values 栏中有两个选项：

- Sequential numbers, 自动赋予顺序数作为索引值。若本例选择此项，索引变量值为 1~3。

- Variable names, 使用所选择的变量组的各变量的变量名作为索引值。从列表中选择一个变量组。本例若选择此项，索引变量值为 scoreA、scoreB、scoreC。

在 Edit the Index Variable Name and Label 表中编辑索引变量属性，即对索引变量，可以改其默认的变量名和输入描述变量的标签。

⑥ 单击“Next”按钮，打开第 6 步对话框，如图 2-71 所示。

- Handling of Variables not Selected 栏，确定如何处理原始数据文件中，未被选择的变量。在选择变量的第三步，选择了要重新构建的变量组和一个当前数据中的标识变量。所选择的变量的数据将出现在新文件中。如果在当前文件中还有其他变量，可以选择丢弃或保留它们。有两个单项。

Drop variable(s) from the new data file, 丢掉未被选择的变量，系统默认。

Keep and treat as fixed variable(s), 保留并作为固定变量处理。

- System Missing or Blank Values in all Transposed Variables 栏，此栏中确定如何处理无效值，即要进行转换的变量中的缺失值和空值，下有两个单项：

Create a case in the new file, 在新文件中生成一个观测量。

Discard the data, 剔除这个数据。

- Case Count Variable 栏，确定是否在转换后的新文件中生成计数变量。计数变量包含当前数据中产生的新行数。如果选择丢弃无效值，计数变量可能是很有用的，因为有可能对给定的当前数据产生不同的行数，只有一个选项。

Count the number of new cases created by the case in the current data, 对由当前数据文件中的一个观测量产生的新观测量的数进行计数。选择此项，可以对计数变量改变默认的变量名和提供描述变量的标签。

⑦ 单击 Next 按钮，打开最后一步对话框，见图 2-72。这是 Restructure Data Wizard 最后一步，有两个选项。

- Restructure the data now, 重新构建数据。将产生新的、重新构建的文件。如果想立刻改变当前数据文件，就选择此项。

**注意：**如果原始数据是被加权的了，新数据也会是加权的，除非用作权重的变量是被重新构建的或者从新文件中去除了。

- Paste the syntax generated by the wizard into a syntax window, wizard 将产生的语句粘贴到语句窗口中。当没有准备好改变当前文件时，或者想修改语句，或者保存它以便以后再用时，选择此项。

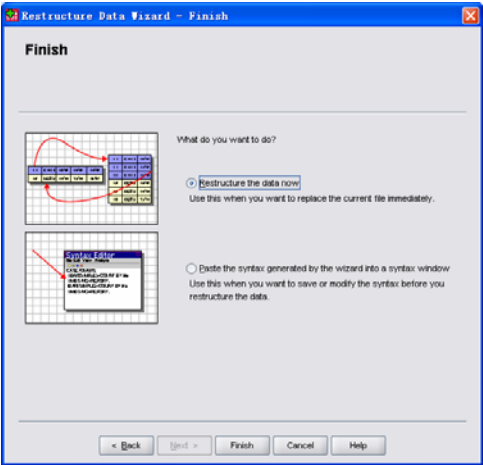
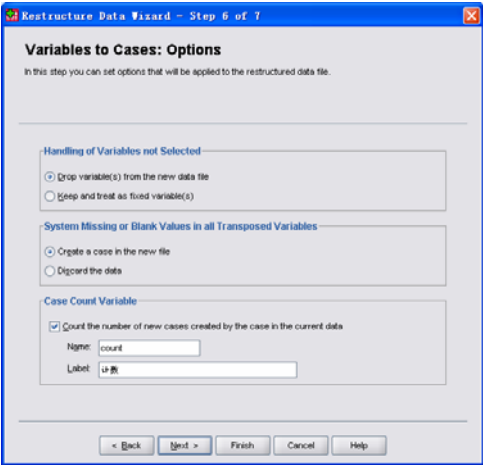


图 2-71 变量组转换到观测量组第 6 步对话框

图 2-72 变量组转换到观测量组第 7 步对话框

选择①，单击“Finish”按钮，程序运行转换，结果如图 2-73 所示。图 2-73(a)原始数据文件；图 2-73(b) 索引变量值选择使用顺序值的结果；图 2-73(c)是保留固定变量  $h$ 、 $w$ ，索引变量使用原始变量名的结果；图 2-73(d)不保留固定变量，索引变量使用原始变量名的结果；。运行结果保存在 data02-13-1a、data02-13-1b 和 data02-13-1c 中。

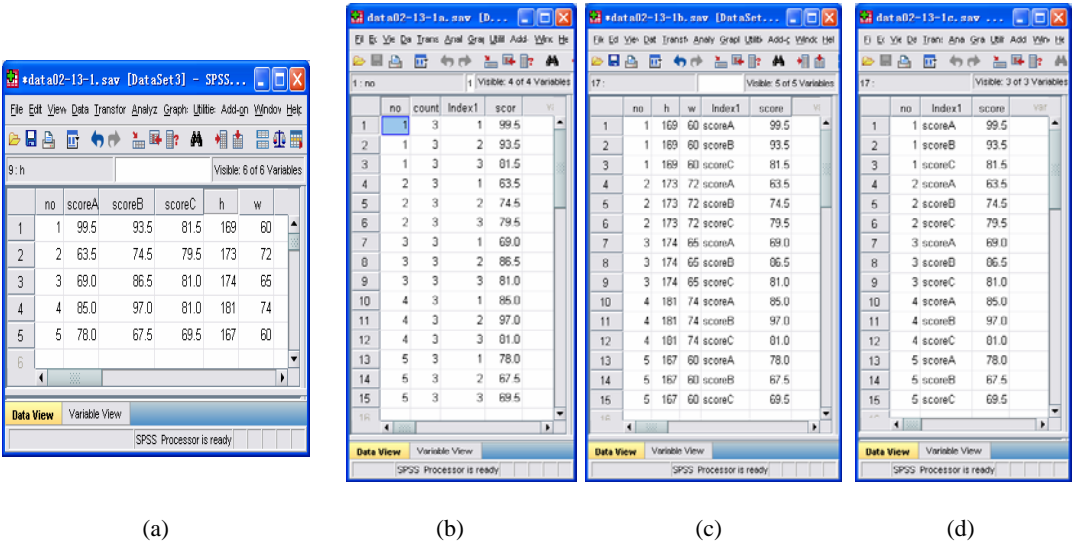


图 2-73 原始数据和不同选项生成的不同新文件

(4) 如果转换两个变量组，数据 data02-13-2。数据包括变量：学号 no、理科成绩 scoreA1、scoreB1、scoreC1 和文科成绩 scoreA2、scoreB2、scoreC2，和身高  $h$ 、体重  $w$ 。

转换结果见图 2-74。转换步骤与上述 7 步基本相同，仅下述操作有区别：

① 在图 2-67 所示的操作第 2 步对话框中选择 **More than one** 项。

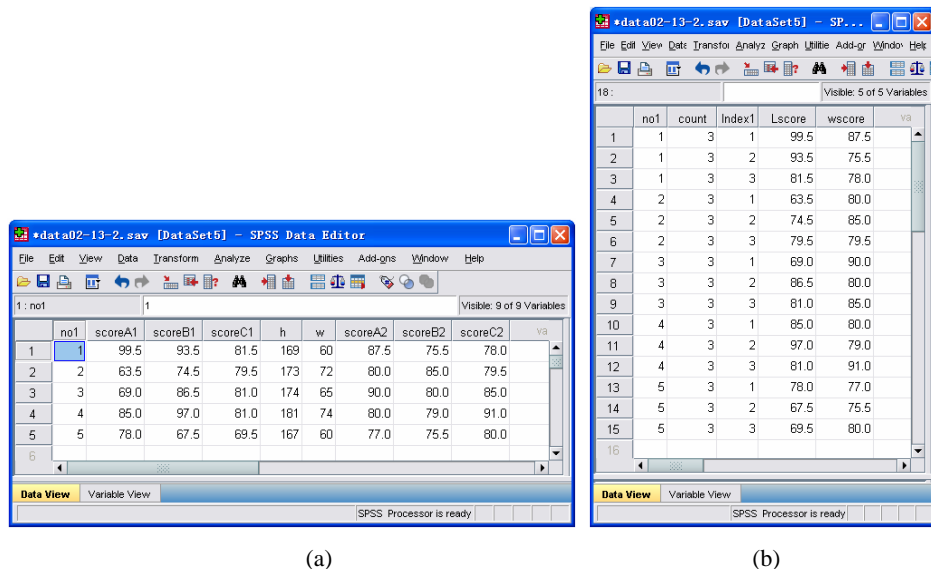


图 2-74 变换两个变量组到观测量的原始数据和新文件数据

② Wizard 第 2 步在图 2-67 所示对话框中，定义了第一组变量 scoreA1、scoreB1、scoreC1，新变量名为 Lscore 后，在 TargetVariable 下拉菜单中选择另一个默认变量名，下面的原始变量栏自动清空。输入新 Target 变量名 Wscore，将变量 scoreA2、scoreB2、scoreC2，送入 Target 变量栏。

③ 第 5 步选择 **Sequential numbers** 项，索引变量值选择为顺序值比较好。其余选项不是必须改变的。原始数据和转换结果如图 2-74 所示。

图 2-74(a)中为原始数据，图 2-74(b)为转换的结果。结果见数据文件 data02-13-2a。

【例 9】两因素不同水平的转换的例题。

(1) 数据 data02-14 是三门课程 A、B、C 不同教材教学的 1、2 两个班的成绩。scoreA1、scoreB1、scoreC1 是 1 班 3 门课程的成绩，scoreA2、scoreB2、scoreC2 是 2 班 3 门课程的成绩。现在进行变量组到观测量的转换。要生成两个索引变量，在分析时作为两个分类变量使用。转换结果见图 2-75。

(2) 主要操作步骤是：

① 在主对话框中，仍选择第一项 **Restructure selected variable into cases** 把选择的变量转换成观测量，见图 2-66。（提示：变量要分类按序存放。）

② 在图 2-67 的第 2 步对话框中选择第 1 项 **One**，转换一组变量成一个新的因变量。



③ 在图 2-68 的第 3 步对话框中, 将 6 个变量 scoreA1、scoreB1、scoreC1、scoreA2、scoreB2、scoreC2 全部送入 Target Variable 下面的变量栏内, 在 Target Variable 栏内输入新变量名 score。

④ 在如图 2-69 的第 4 步对话框中, 选择第 2 项 More than one, 建立不止一个索引变量。也就是说有所选择的一组变量属于不止一个因素 (条件或处理) 的因变量。

⑤ 在如图 2-71(b)所示的第 5 步对话框中, 在表中填写变量名、变量标签和水平值。将变量名 Index1 改为 class、同行的 Label 单元输入 “班级”、同行的 Levels 单元输入水平数 2; 在下一行各单元格分别输入: courses、课程、3。(提示: 此处的班级和课程顺序是对应原数据中的变量顺序的, 在原数据中变量首先按班级排序, 然后按课程排序, 所以在此处第一个索引变量是班级, 第二个索引变量是课程, 注意要对应, 否则结果将不是操作者所想要的。)

⑥ 最后一步如果选择第 2 项, 在语句框中生成的过程语句为:

```
VARSTOCASES /ID = id
```

```
/MAKE score FROM scoreA1 scoreB1 scoreC1 scoreA2 scoreB2 scoreC2
```

```
/INDEX = class "班级"(2) courses "课程"(3) /KEEP = /NULL = KEEP.
```

修改/MAKE 子命令成: /MAKE score "成绩" FROM scoreA1 scoreB1 scoreC1 scoreA2 scoreB2 scoreC2 并运行之, 因为对话框中没有空间输入新变量标签。如果不做如上修改, 新变量标签与第一个原始变量的标签相同。运行结果见图 2-75, (a)图为原始数据文件, (b)图为转换的第 5 步, (c)图为新数据文件的一部分, 详见 data02-14a。

(3) 变量组转换成观测量的过程语句

VARSTOCASES	①
/MAKE new variable ["label"] [FROM] varlist [/MAKE ...]	②
/INDEX = {new variable ["label"]} {new variable ["label"] (make variable name)}	
{new variable["label"] (n) new variable ["label"] (n)}	③
/ID = new variable ["label"]	④
/NULL = {DROP**} {KEEP }	⑤
/COUNT=new variable ["label"]	⑥
/KEEP={ALL** } {varlist}	⑦
/DROP=varlist	⑧

① 变量组转换成观测量调用 VARSTOCASES 过程。VARSTOCASES 是调用语句关键字, 必须在程序首行。

② MAKE 子命令可以有若干个, 每个定义一个新变量和原始变量组。关键字后是新变量名和用双引号中的变量标签, 紧接着是构成新变量的原始变量表。

③ INDEX 子命令等号后面是定义的索引变量名和双引号中的标签, 如果使用顺序值做索引变量值, 标签后写上顺序最大值 n, 也就是 n 个水平; 如果使用变量组的变量名

做索引变量的值, 则后面写上 **make** 子命令中的原始变量表。除命令关键字外, 都是可以选择的内容。

④ **ID** 子命令定义标识变量, 在关键字后面的等号连接标识变量名和变量标签。

⑤ **NULL** 子命令确定如何处理无效值, 默认的是 **DROP**, 即如果原始变量值无效, 新变量中去掉这个观测量; 选项 **KEEP** 是保留这个无效值的观测量。

(a)

Index Variable Names, Labels, and Number of Levels:	Name	Label	Levels	Index Values
1 class	课程		2, 3	
2 courses	课程		1, 2, 3	

Total number of combined levels (product): 6

Note that the product of the number of levels of all the index variables must equal the "Total" number displayed below the table. This number equals the number of variables to be transposed.

(b)

	id	class	courses	score
1	1	1	1	99.5
2	1	1	2	93.5
3	1	1	3	81.5
4	1	2	1	87.5
5	1	2	2	75.5
6	1	2	3	78.0
7	2	1	1	63.5
8	2	1	2	74.5
9	2	1	3	79.5
10	2	2	1	80.0
11	2	2	2	85.0
12	2	2	3	79.5
13	3	1	1	69.0
14	3	1	2	86.5
15	3	1	3	81.0
16	3	2	1	90.0
17	3	2	2	80.0
18	3	2	3	85.0
19	4	1	1	85.0
20	4	1	2	97.0
21	4	1	3	81.0
22	4	2	1	80.0
23	4	2	2	79.0
24	4	2	3	91.0
25	5	1	1	78.0
26	5	1	2	67.5

(c)

图 2-75 原数据文件、转换的第 5 步对话框和部分新文件

⑥ **COUNT** 子命令定义计数变量。关键字后等号连接计数变量的变量名和标签。

⑦、⑧ **KEEP** (或 **DROP**) 子命令确定除要转换的变量组外的固定变量, 在新数据文件中, 是保留 (还是去掉), 关键字后等号连接要保留 (或去掉) 的变量表。

### 【例 10】观测量组到变量组结构的转换。

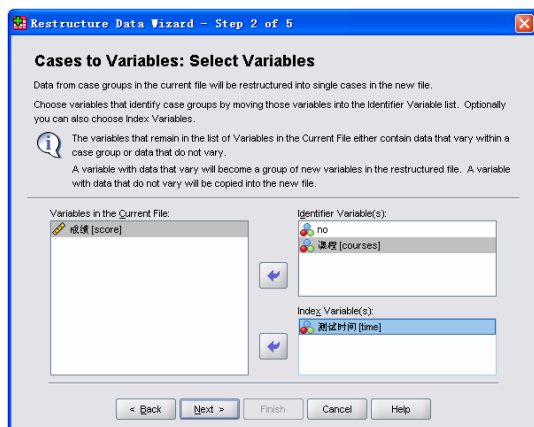
以 data02-15 为例说明转换操作。变量 **time** 为 2 水平的因素, 表示两个测试时间, 期中和期末; **courses** 是 3 水平的因素, 表示 3 门课程 A、B、C。score 是分别对 5 个学生的测试成绩。这是应该进行一个因素即 3 门课程, 两次 (期中、期末) 重复测量的方差分析的问题。而这个数据结构不符合分析要求, 必须转换。**id** 是接受测试的 5 个学生的标识变量。要按 **time** 变量的两个水平将分数 **score** 变成两个变量表示期中和期末对 3

门课程的测试分数, 操作如下。

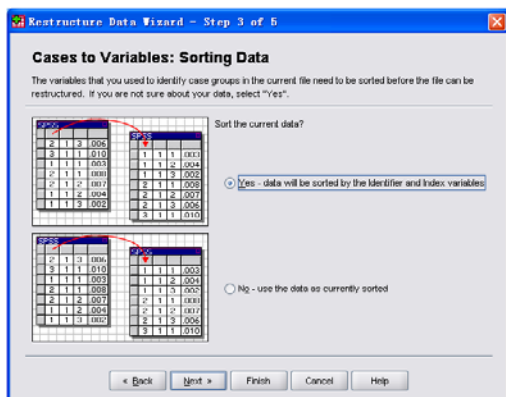
(1) 在 Restructure Data Wizard 对话框中, 见图 2-66, 选择第 2 项 Restructure selected cases into variables 把选择的观测量组变成变量组。单击 Next 按钮。

(2) 打开如图 2-76(a)的第 2 步对话框。将新文件中作为分类变量的 courses (课程) 和标识变量的 no (学生编号) 送入 Identifier Variable(s) 栏中; 将要按其转换的分类变量 time 送入 Index Variable(s) 栏内。单击 Next 按钮。

(3) 打开如图 2-76(b)的第 3 步对话框。有两个选项, 确定是否对原始数据文件按标识变量排序。第 1 项 Yes, 表明要 Wizard 对原始数据文件按前一步指定的标识变量排序。如果没有排序或不确定是否已经排好序, 应该选择此项; 第 2 项 No, 当原始数据文件已经按标识变量排好序了, 选择此项。Wizard 每遇到标识变量值的一个新的组合, 就生成一个新行, 所以, 对数据文件按标识观测量组的变量值排序很重要的。选择后单击 Next 按钮。



(a)



(b)

图 2-76 观测量转换到变量组第 2 步和第 3 步

(4) 打开如图 2-77 所示的第 4 步对话框, 共有三类选项。

① Order of New Variable Groups 栏, 确定新变量组的顺序, 对要转换成两组以上新变量时选择才有意义。我们的例题因只生成两个新变量, 不需要选择。此类选项, 有两种排列方式:

Group by original variable, 按原始变量顺序成组排列, 例如 (w1、w2、w3、h1、h2、h3)。

Group by index, 按索引变量值转换排序。例如 (w1、h1、w2、h2、w3、h3)。

② Case Count Variable 栏, 确定是否生成计数变量。选中 Count the number of cases in the current data used to create a new case, 在 Name 和 Label 框中分别给出变量名和标签。

③ Indicator Variables 栏, 确定是否生成指针变量。选中 Create indicator variables,

在 Root Name 框中给出变量名字头。

Wizard 可以用索引变量在新文件中生成指针变量。对索引变量的每个值生成一个新变量。指针变量指明观测的一个值是出现与否。如果观测量有值，指针变量的值是 1，否则值为 0。在某些问题中，指针变量可以做频数计数用。对本例题没有用，所以不选。

④ 最后一步确定是立即执行，还是先生成过程语句，选择第 2 项。数据转换前后的结果见图 2-78。

过程语句如下：

`SORT CASES BY id courses time .`

`CASESTOVARS /ID = id courses /INDEX = time /GROUPBY = VARIABLE .`

`SORT CASES` 命令按 `by` 后面的变量表对原始数据文件排序。

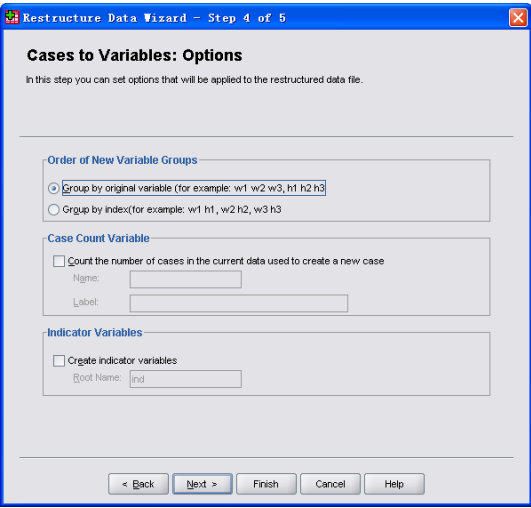


图 2-77 观测量组转换成变量组的第 4 步对话框

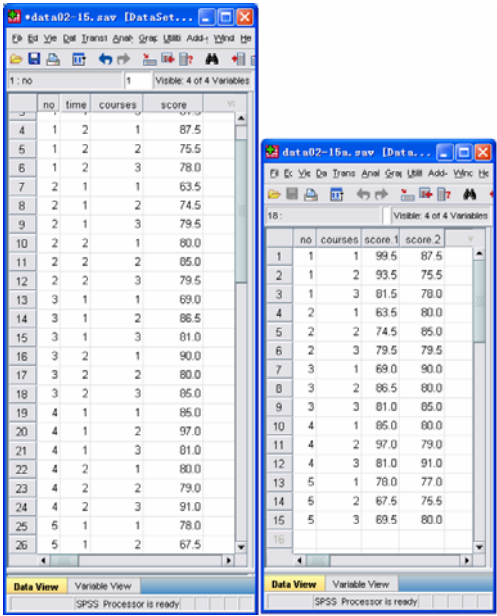


图 2-78 转换前后的部分数据

`CASESTOVARS` 命令对原始数据文件进行观测量组到变量组的转换；`ID` 子命令指定新文件中的分类变量，`INDEX` 子命令指定数据按等号后面的变量水平生成变量组。有几个水平，生成的变量组就包括几个变量；`GROUPBY` 子命令要求按变量成组排列新生成的变量。

(5) 观测量组转换成变量组的过程语句

- CASESTOVARS [ /ID = varlist] [ /FIXED = varlist]
- [ /VIND [ROOT = rootname]]
- [ /COUNT = new variable ["label"]]
- ①
- ②
- ③

[/RENAME varname=rootname varname=rootname ...]	④
[/SEPARATOR = {"." } {"string"}]]	⑤
[/INDEX = varlist]	⑥
[/GROUPBY = {VARIABLE**} {INDEX }]]	⑦
[/DROP = varlist]	⑧

① CASESTOVARS 命令 调用观测量组到变量组转换的过程；ID 子命令用等号连接标识变量表；FIXED 子命令指定带到新文件中的固定变量，即不进行转换的变量，等号连接变量表。

② VIND 子命令 指定生成指针变量，选项 ROOT 用等号连接指定的变量名字头。

③ COUNT 子命令 指定生成计数变量，在等号后面给出新文件中计数变量的变量名和变量标签。

④ RENAME 子命令 定义转换后变量名的基本部分。新变量名的序号是原始分类变量的水平序号。

⑤ SEPARATOR 子命令 定义分隔符，默认的分隔符是英文半角的圆点。

⑥ INDEX 子命令 在等号后面指定在新文件中索引变量，等号后面给出索引变量表。

⑦ GROUPBY 子命令 指定新变量排列顺序，选项 VARIABLE 是默认的，是一个排列在一起的变量组。选项 INDEX 是按索引变量值排列的。

⑧ DROP 子命令 指定在原始数据文件中的非转换变量是否在新文件中出现。要去掉的变量与选项关键字用等号连接。

## 2.4 观测量的加权与选择

### 2.4.1 定义加权变量

在实际应用中，我们经常需要对观测量进行加权处理。例如，在数据文件中如果存在一个表明相同的变量值出现频数的变量时，应该定义该变量为权重变量。可以选择 Data 菜单中的 Weight Cases 命令，定义权重变量。至于对哪个变量的值加权，是使用权重变量计算中的问题。

1. 在选择加权变量时应该注意以下三点：

- (1) 权重变量中含有零、负数或缺失值的观测量将被排除在分析之外；
- (2) 分数权重值有效；
- (3) 一旦定义了权重变量，那么在以后的分析中权重变量一直有效，直到取消权重变量的定义，或者定义了其他的权重变量。

2. 定义权重变量的操作如下。

- (1) 按 Data→Weight Cases 顺序打开 Weight Cases 对话框，如图 2-79 所示。

(2) 选择是否对观测量进行加权处理。

① Do not weight cases 是系统默认状态, 表示对数据不加权, 不用定义权重变量。

② Weight cases by, 选择此项要求对观测量加权。

(3) 选择加权变量。从左边源变量框中选择权重变量, 送入 Frequency Variable 框中。

(4) 单击 OK 图标按钮, 权重变量定义完成。

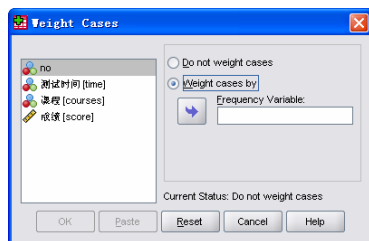


图 2-79 定义加权变量

## 2.4.2 选择参与分析的观测量

如果需要部分观测量参与分析, 就要在分析之前进行选择, 操作方法如下。

1. 单击 Data→Select Cases 打开对话框, 如图 2-80 所示。

2. 几种选择方法:

(1) All cases 选项是系统默认的, 全部观测量都参与分析, 不做选择。

(2) If condition is satisfied, 选择满足条件的观测量。单击 If 按钮打开图 2-81 对话框, 设置选择条件。例如只选择女性, 原变量表中选择变量 gender 送入条件编辑栏, 输入“=“F””, 单击 Continue 按钮。

(3) Random sample of cases 选项, 对数据文件中的观测量进行随机采样。单击 Sample 按钮, 打开如图 2-82 的二级对话框。有两种采样方法, 选择其中一种。

- 按给定的百分比近似选择。在 Approximately 后面输入百分比数值。

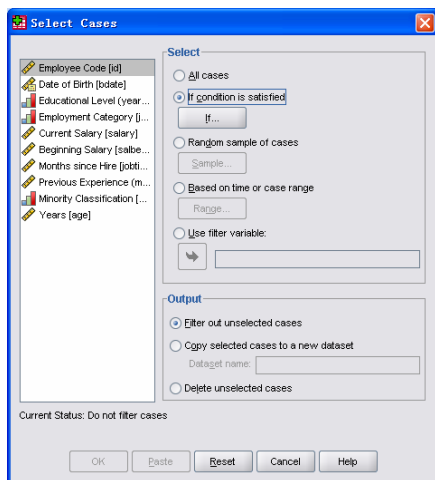


图 2-80 选择观测量主对话框

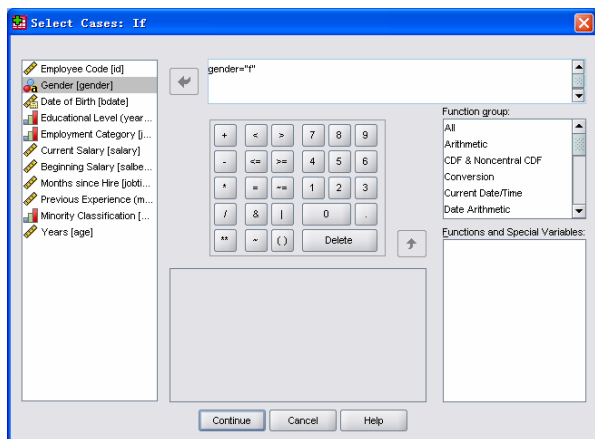


图 2-81 设置选择条件对话框

- 在指定范围内随机选择给定数目的观测量。Exactly 后面输入样本量  $n1$ , 在 cases from the first 后面输入一个小于或等于全部观测量数的数值  $n2$ 。选择观测量是从前  $n2$  个

观测量中选择出  $n1$  个观测量。

此种方法属于重复采样，一个观测量可能被选中不只一次，因此样本量与全部观测量之比是近似等于给定的百分比，或近似等于指定的样本量数值。

(4) Based on time or case range 选项，根据时间或数据范围选择，打开如图 2-83 的对话框，输入第一个观测量号和最后一个观测量号，给定范围之内的观测量被选中。

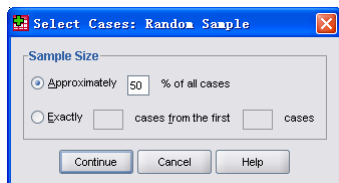


图 2-82 随机采样

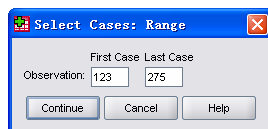


图 2-83 按范围选择

(5) Use filter variable 选项，使用过滤变量，从左面的源变量栏中选择一个数值型变量作为过滤变量，过滤变量值不是 0，或不是缺失值的观测量都被选中。

以上 5 种选择方法是单选项，选择一种，设置好条件参数，返回主对话框。

(6) 在 output 栏中，选择未被选中的观测量的处理方法。

- Filter out unselected cases，观测量号被打上斜线，不参与分析。这是系统默认的处理方法。

- Copy selected cases to a new dataset 把被选中的观测量复制到一个新数据集中。选择此项，在 Dataset name 栏中输入新文件名。

- Deleted unselected cases，未被选中的从数据文件中删除。

参数设置完成，单击 OK 按钮，选择完成。

## 习 题 2

1. SPSS 的变量有几种类型？
2. 变量的哪些属性会影响它们在分析中的作用？哪些属性只影响数据在窗口中的显示？哪些属性只影响输出？
3. 变量有哪几种测度方式？在分析中的作用是什么？
4. 你的工作中数据存放在什么格式的文件中？SPSS 可以直接打开这些数据文件吗？如果不能直接打开，是否能经过转换形成 SPSS 格式的数据文件？
5. 数据文件 data02-01 中，用查重功能分析是否受教育年限相同的职工，初始工资都相同。
6. 为什么要拆分数据文件？拆分的结果是什么？
7. 合并数据文件有几种情况？

8. 观测量排序和排秩有什么区别？什么叫结？结上观测量的秩次有几种排法？体育比赛常用哪种方法排列名次？

9. 查看 data02-03.txt，确定它是否是固定格式的 ASCII 码数据文件？有列间隔吗？将其转换为 SPSS 数据文件。

10. 将 data02-01 数据文件按 educ 变量值升序排列。

11. 为什么要对变量重新编码？SPSS 有几种重新编码的过程？举例说明。

12. 什么是数据文件的重新构建？有几种重新构建的方式？

13. 超市对竞争对手的商品价格做定期调查。某天，某超市调查了 3 个竞争超市 49 种商品的售价，与本超市进行比较。从 49 种中随机抽取 7 种商品的价格。数据记录在 data02-16 中。因要做方差分析，要求将 4 个超市的商品价格放到一个价格变量中，另外增加变量，使数据文件能正确表达每个价格是属于哪个超市哪个商品的。



## 第 3 章 输出信息的编辑

如果文本窗口中默认的常用工具不全,操作不方便。最好使用 View 菜单中的 Toolbars 功能将常用工具按钮显示在工具栏中。见第 1 章 1.2.6 节内容。例如在默认工具栏中加入剪切、复制、粘贴、删除等常用图标按钮。

SPSS 使用与 Windows 系统相同的基本编辑功能和图标按钮。查找与替换的操作与 Windows 系统的同类功能一致。本章只介绍语句窗和输出窗口中的一些特殊的常用编辑方法。

### 3.1 输出窗口中的文本浏览与编辑

系统中的操作与过程运行的结果显示在输出窗口 Viewer 窗口中。输出窗口的导航系统是比较特殊的输出窗口信息浏览器,它不但为窗口中内容查找、浏览提供了工具,同时也为窗口中的内容编辑提供了方便。

#### 3.1.1 利用导航器浏览输出信息

图 3-1 是输出窗口 Viewer。左半部分是导航器,右半部分是输出信息区。导航器实际上是可折叠的输出信息树形结构图。对导航器中的每一项都可以使用鼠标左键单击进行选择;双击进行打开与隐藏的操作。

##### 1. 认识导航器

导航器中有 Output 输出总项,这是最高一级输出项;以过程语句命名的过程项,如图 3-1 中的 Frequencies、Descriptives 过程项是第二级;第一、二级输出项前显示的是带有结构图的书形图标,而且有折叠图标(加减号)。每个过程项都可能

包括几种结构项,即第三级结构项。这些结构项是否显示在导航器中取决于系统参数设置,见 1.3 节有关内容。可能显示的结构项有:Log 日志项、Title 标题项、Notes 说明项、

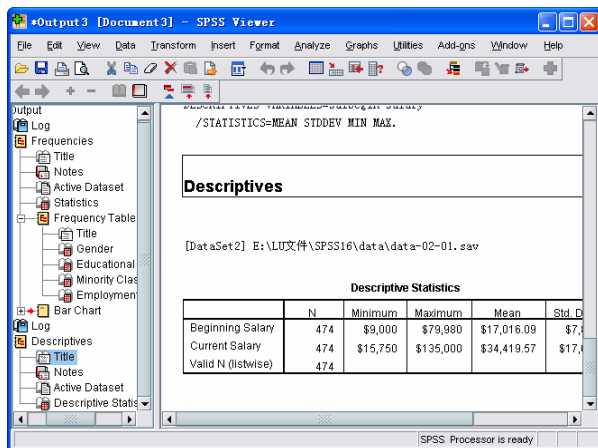


图 3-1 Viewer 输出窗口

Pivot Table 输出表格项、Warnings 警告项、Chart 统计图项，如图 3-1 中的 Bar Chart 和 Text Output 文本输出项。

## 2. 在导航器中选择输出项

单击在导航器结构图中的某一项，与该项相应的输出信息显示在右侧窗口的可见部位，且外轮廓加了黑实线框。用这种方法可以将需要浏览的部分调入窗口中可见信息区。

## 3. 在导航器中关闭/打开输出项

为突出浏览重点，可隐藏部分输出线，需要浏览时再显示出来，操作如下。

(1) 导航窗口中一级控制项是 **Output**，单击前面的加（或减）号图标，所有输出内容全部显示（或隐藏），见图 3-2(a)。

(2) 隐藏/显示各级内容。导航窗口中第二级是过程控制。每一项是一个过程输出。单击过程项前面的加（减）号图标，可以显示（或隐藏）过程项中的内容。如图 3-2(a)、图 3-2(b)分别是二级项隐藏和打开的状态。单击书形图标，所有第三级项内容在右侧窗口中显示或隐藏。同时相应的输出信息也在信息区中显示或消失，如图 3-2(c)的 Descriptives 是被隐藏了第三级结构项的过程项。而 Frequencies 过程项是被打开的。

(3) 显示某项输出在右面信息窗口中的方法是在导航器中单击三级项，该项前面出现红色箭头，书形图标打开，相应内容在右侧窗口可见。在对应的信息（一个标题、一个表格、一个统计图或一段完整的文本）左面也显示出红色箭头，外边显示黑色框线。

如果第 3 级不只一个输出项，例如图 3-2(c)中的条形图有 4 个，鼠标单击一项，右侧窗口显示相应的输出项。

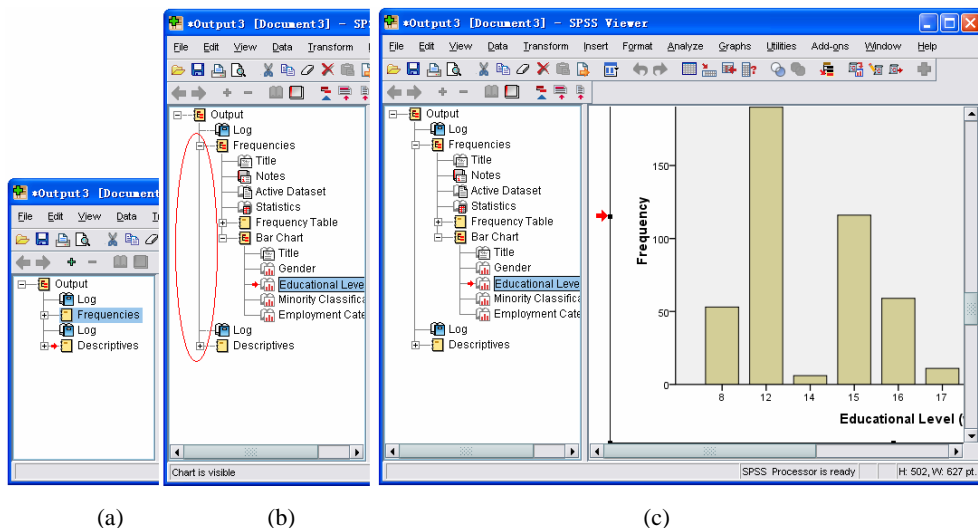


图 3-2 输出项的打开与隐藏


另外，使用位于导航器的上方工具栏中的图标按钮也可以打开或关闭任意一个结构项。方法是鼠标左键单击选定的结构项，使之彩底显示。单击工具栏中合着或打开的可

以显示或隐藏选择的结构项; 导航器上方的向左/右箭头图标按钮也能完成上述操作。读者可以自己操作。打开需要显示的内容, 隐藏暂不需要显示的内容, 会使观察、编辑和选用输出信息变得方便。

### 3.1.2 编辑导航器中的输出项

#### 1. 选择操作对象

使用 Edit 菜单项 Select 子菜单中的选择功能对操作对象进行分类选择, 见图 3-3。子菜单中的各项选择功能如下。

- Last output, 即选择最后一个输出项。指最后一次执行 SPSS 过程的全部输出。本功能相应的图标按钮是  Select Last Output。

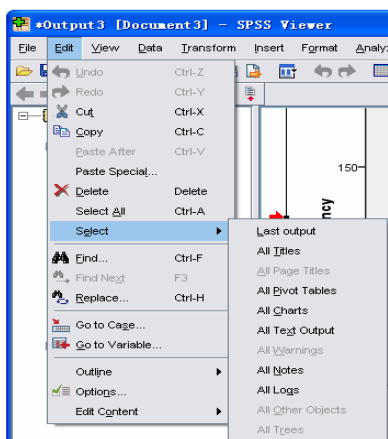


图 3-3 Edit 菜单和 Select 子菜单

- All Titles, 选择所有输出的标题。
- All Pivot Tables, 选择所有输出表格。
- All Page Titles 选择所有页面标题。
- All Charts, 选择所有统计图。
- All Text Output, 选择所有文字信息。
- All Warnings, 选择所有警告信息。
- All Notes, 选择所有说明信息。
- All Logs, 选择所有日志信息。
- All Other Objects, 选择所有其他对象, 也就是上述各项未包括的信息。
- All Tree, 选择所有树形图。

#### 2. 使用鼠标键选择操作对象

(1) 可以用鼠标左键在导航器中单击一个结

构项选择一个操作对象, 使之彩底显示。

(2) 按住 Ctrl 键的同时鼠标左键单击要选择的对象, 可选择位置不连续的多个对象。

(3) 按住 Shift 键的同时用鼠标左键分别单击两个不相邻的结构项, 可以选择这两个结构项之间的各项。

3. 对被选中的结构项及其内容可以进行删除、剪切到剪贴板、复制到剪贴板和粘贴到另一位置的操作。

4. 移动输出项显示位置的另一种方法是用鼠标拖动到目标位置, 见图 3-4。用这种方法可以将有用信息组织到一起。

如果在按下鼠标左键的同时按下 Ctrl 键, 同样的操作会复制选定的输出项。

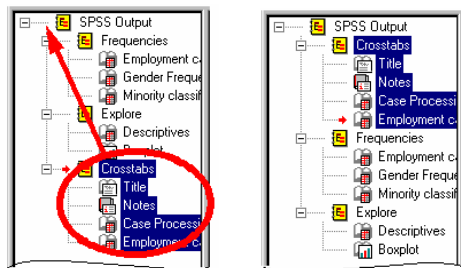


图 3-4 使用鼠标拖曳移动输出项

用这种方法可以重新组织输出信息的顺序。虽然各级复制项都可以使用鼠标拖曳到任何新位置,但是把一个输出项拖到另一个过程项中只能使输出信息混乱。建议只在一个过程项内做这种操作。

## 3.2 输出表格中信息的编辑

### 3.2.1 表格编辑工具与常用编辑方法

#### 1. 选择操作对象

双击要编辑的表格或过程输出标题即选择了这个表格或标题。要在表格中具体选择表格元素,使用下面的方法:

- (1) 选择一行,按住 Ctrl+Alt 组合键,用鼠标单击表格左端的行栏目,见图 3-5(a);
- (2) 选择一列,按住 Ctrl+Alt 组合键,用鼠标单击表格的列栏目见图 3-5(b);
- (3) 选择不只一行或一列,可以按住鼠标拖动,所经过的行(或列)均反向显示;
- (4) 在有的表中,选择一行(或一列)会导致把与之相关的行(或列)也同时被选择,如图 3-5(c)。

ANOVA					
Dependent Variable	Current Salary				
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8.94E+10	2	4.472E+10	434.5	.000
Within Groups	4.85E+10	471	102925714.5		
Total	1.38E+11	473			

(a)

ANOVA					
Dependent Variable	Current Salary				
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8.94E+10	2	4.472E+10	434.5	.000
Within Groups	4.85E+10	471	102925714.5		
Total	1.38E+11	473			

(b)

Multiple Comparisons						
Dependent Variable		Final Height				
		Statistics				
Test	(I) Fertilizer	(J) Fertilizer	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound Upper Bound
Tukey HSD	1	2	-5.00	3.48	.407	-17.39 7.39
		3	-16.00*	3.48	.022	-28.39 -3.61
	2	1	5.00	3.48	.407	-7.39 17.39
		3	-11.00	3.81	.093	-24.57 2.57
	3	1	16.00*	3.48	.022	3.61 28.39
		2	11.00	3.48	.407	-7.39 17.39

Based on observed means.  
\*. The mean difference is significant at the .05 level.

(c)


图 3-5 选择表中的操作对象

2. 表格编辑工具

双击表格，会显示表格编辑工具栏，如图 3-6 所示。



图 3-6 表格编辑工具栏

表格编辑工具栏工具及功能除  Pivot Controls 表格控制功能外，其他功能如撤销与恢复操作、单元格中字体字号的设置、对齐方式的设置都与 Windows 中相应功能的操作方法相同。

3. 标题编辑工具

双击一个标题，会打开标题编辑工具栏。如图 3-7 所示，上面是被编辑的标题，下面是工具栏。具体选择了标题区中的文字，即可进行编辑，包括修改文字内容、改变字体、字号、对齐方式等。

4. 修改单元格中的内容

修改表格单元格中任何内容均可以仿照修改栏目名称的操作方法，但最好不要修改单元格中的数据。除了应该实事求是外，还因为修改一个数据会影响其他数据的正确性，这是 SPSS 的输出与 Excel 工作表的不同之处。例如在图 3-8 中，将变量 SALARY 的 Minimum 值由原来的\$15,750 改为\$1,575，逗号分隔符位置改变引起数值变化，与之联系的 Range 值不会自动随之改变，以致表中 Range 值是错误的。

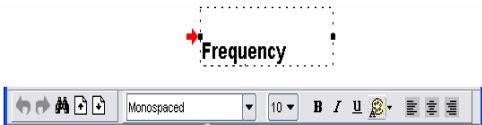
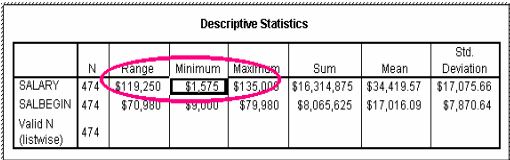


图 3-7 标题及标题编辑工具栏



	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation
SALARY	474	\$19,250	\$1,575	\$135,000	\$16,314,875	\$34,419.57	\$17,075.66
SALBEGIN	474	\$70,980	\$9,000	\$79,980	\$8,065,625	\$17,016.09	\$7,870.64
Valid N (listwise)	474						

图 3-8 修改表格中的数据会造成错误

5. 隐藏或显示表格的行与列

(1) 双击选定的表格，选择要隐藏的行或列。

(2) 单击 View 菜单，选择 Hide 命令，选定的表格或栏目被隐藏起来。

在有的表中，选择一行（或一列）会导致把与之相关的行（或列）也同时被选择，如图 3-5(c)所示，因此需要隐藏时必须将同一类都隐藏起来，否则表中数据无法解释。

(3) 如果想恢复显示被隐藏的行或列，必须先选择邻近的未隐藏的行或列，再选择 View 主菜单中的 Show 命令。

6. 改变表格列宽度

表格的显示常常因列宽不够将数字显示成一系列星号，这就需要调整表格的列宽。

### (1) 手动调整

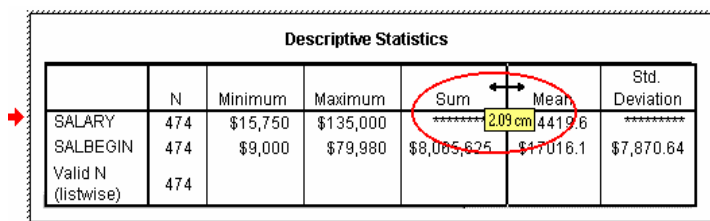
将鼠标光标置于要调整的表格竖线上，此时光标变为水平的双箭头线，同时待调整的竖线加粗。按住鼠标左键，拖动鼠标调整表格列宽，直到列宽合适，或未显示的数值显示出来为止，松开鼠标左键，见图 3-9。

应该注意，调整列宽时显示的列宽数值，可作为调整参考。列宽的调整会影响列中数据显示的有效数字位数。当调整列宽过小时，会显示出“Hide”字样，如果此时松开鼠标左键，列宽过小的栏目会被隐藏。如果发现因宽度调整而隐藏了一列不应隐藏的数据，可以使用 Edit 菜单中的 Undo 命令将其撤销。

### (2) 菜单命令调整

双击选择表格，按 Format→Set Data Cell Width 顺序单击菜单项，打开如图 3-10 所示的调整列宽度对话框。在 Width for all data cells 后面的编辑区中输入宽度值，也可以单击向上、向下箭头按钮增加或减少宽度数值。单击 OK 按钮确认。

这样设置的宽度产生的效果是除最左面一列外，其他各列等宽。



	N	Minimum	Maximum	Sum	Mean	Std. Deviation
SALARY	474	\$15,750	\$135,000	*****2.09 cm	4419.6	*****
SALBEGIN	474	\$9,000	\$79,980	\$8,005,625	\$17016.1	\$7,870.64
Valid N (listwise)	474					

图 3-9 手动调整表格的列宽

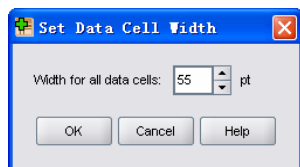


图 3-10 调整列宽度对话框

## 3.2.2 表格的转置与行、列、层的处理

表格是运行分析过程自动产生的。表格形式不一定能满足编写报告的要求，例如行、列的安排使得表格过长或过宽，都会在一定程度上影响对数据的观察和分析。可以用下面的方法进行转换。

### 1. 使用菜单对表格进行行、列互换（转置）

(1) 双击选定的表格，使之显示外阴影框。此时输出窗口中的主菜单发生改变。

(2) 按 Pivot→Transpose Rows and Columns 顺序单击菜单项。表格完成转置。图 3-11(a) 是对一个频数分布交叉表进行转置，转置后的结果如图 3-11(b)所示。

### 2. 使用表格转置盘进行行、列、层之间的位置转换

在输出信息区，双击要进行编辑的表格，再按 Pivot→Pivoting Trays 顺序单击菜单项。都会显示表格转置工具盘，如图 3-12 所示，盘标题栏标有 Pivoting Trays。

转置盘有 3 个图标，分别代表层、行、列。左边标有 Layer 的是表的“层标”，下边与 Row 并列的是“行标”，在转置盘右上方标有 Column 处的是“列标”。如果信息区中选择的是二维表格，则只有行标和列标。

3. 利用转置盘变换层、行、列的位置

使用鼠标拖曳层标、行标、列标中的任意一个到另一个位置，可以改变层、行、列的相互关系，使层、行、列上的数据转换位置，使表格满足显示要求。

图 3-12(a) 是单因变量身高，按性别、年龄的组合分组的均值比较表。最左面一列又分两列（无竖线），分别是年龄和性别。该列中每个单元格是一个组合。每个组合的统计量分别为 Mean、N、Std Deviation 各列对应的单元格中。

Educational Level (years) \* Gender Crosstabulation

Count		Gender		Total
		Female	Male	
Educational Level (years)	8	30	23	53
	12	128	62	190
	14	0	6	6
	15	33	83	116
	16	24	35	59
	17	1	10	11
	18	0	9	9
	19	0	27	27
	20	0	2	2
	21	0	1	1
Total		216	258	474

(a)

Educational Level (years) \* Gender Crosstabulation

Count		Educational Level (years)										Total
		8	12	14	15	16	17	18	19	20	21	
Gender	Female	30	128	0	33	24	1	0	0	0	0	216
	Male	23	62	6	83	35	10	9	27	2	1	258
Total		53	190	6	116	59	11	9	27	2	1	474

(b)

图 3-11 转置前、后的表格

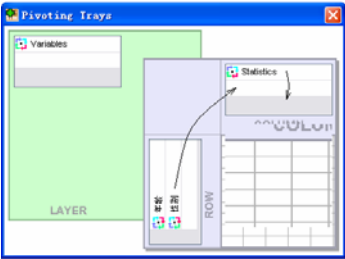
双击表格，单击 Pivot→Pivoting Trays 打开转盘如图 3-12(b)所示。层 Layer 上标有 Variable 变量；行 Row 上标有两个分类变量：性别和年龄。列 Column 上标有 Statistics 统计量；这个表太大，分类变量组合表现也不明显。所以可以把性别变量放到列上，拖曳性别标到 Column 区，然后把原来列上的 Statistics 向下拖曳，使它出现在性别分组的下面。见图 3-12(b)。

拖曳行、列标的同时表格变化。转盘上的标拖曳结果为 3-12(c)。变化结果为图 3-13。

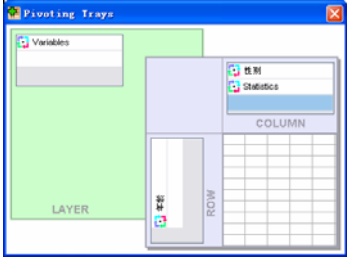
Report

身高	性别	Mean	N	Std. Deviation
10	女	1.4500	5	.02000
	男	1.4467	3	.02887
	Total	1.4488	8	.02167
11	女	1.5393	6	.02317
	男	1.5000	5	.04637
	Total	1.5209	11	.03910
12	女	1.6100	2	.01414
	男	1.6140	5	.01949
	Total	1.6129	7	.01704
13	男	1.5900	1	.
	女	1.5900	1	.
	Total	1.5154	13	.06253
Total	女	1.5357	14	.07623
	男	1.5259	27	.06941
	Total			

(a)




(b)



(c)

图 3-12 输出信息区中的表格及转盘对表格转置的操作

比较图 3-12(a)中转置前的表格和图 3-13 转置后的表格，再比较两图转置盘中图标的位置变化，可以看到效果。

4. 层变量位置的变换

为了便于观察，层变量一般放在表格左上角，如果层变量有两个以上类别，会形成下拉列表形式，见图 3-14(a)。作为报告的一部分，下拉列表不能起作用，因此可以把层变量拖曳到行或列上。层变量放到列上或行上，会使表格变宽或变高。但这两种表格结构可以清晰地显示所有输出结果。图 3-14(b)是图 3-14(a)的转置盘，将层 LAYER 上的变量性别拖曳到行 ROW 上的转置盘为图 3-14(c)所示。将层变量变化到行上的结果见图 3-14(d)。

相反的转变，把在行或列上的分类变量拖曳到层上生成层，对观察输出结果也有很大好处。

Report

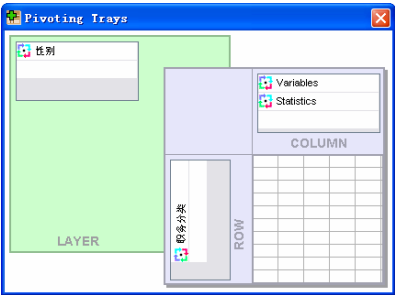
身高	女			男			Total		
	Mean	N	Std. Deviation	Mean	N	Std. Deviation	Mean	N	Std. Deviation
10	1.45E0	5	.02000	1.44E0	3	.02887	1.44E0	8	.02167
11	1.53E0	6	.02317	1.50E0	5	.04637	1.52E0	11	.03910
12	1.61E0	2	.01414	1.61E0	5	.01949	1.61E0	7	.01704
13				1.59E0	1		1.59E0	1	
Total	1.51E0	13	.06253	1.53E0	14	.07623	1.52E0	27	.06941

图 3-13 转换后的表格

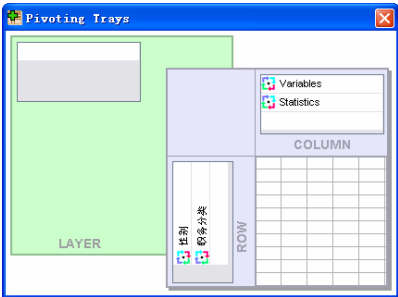
Report

性别	Variables		
	当前工资		
Total	Mean	N	Std. Deviation
办事员	\$31,558.15	157	\$7,997.978
保管员	\$30,938.89	27	\$2,114.616
经理	\$66,243.24	74	\$18,051.570
Total	\$41,441.78	258	\$19,499.214

(a)



(b)



(c)

Report

性别	Variables		
	当前工资		
办事员	Mean	N	Std. Deviation
女	\$25,003.69	206	\$5,812.838
经理	\$47,213.50	10	\$8,501.253
Total	\$26,031.92	216	\$7,558.021
男	\$31,558.15	157	\$7,997.978
保管员	\$30,938.89	27	\$2,114.616
经理	\$66,243.24	74	\$18,051.570
Total	\$41,441.78	258	\$19,499.214
Total	\$27,838.54	363	\$7,567.995
保管员	\$30,938.89	27	\$2,114.616
经理	\$63,977.80	84	\$18,244.776
Total	\$34,419.57	474	\$17,075.661

(d)

图 3-14 将层变量转换为行变量的过程



### 3.2.3 表格外观的设置与编辑

#### 1. 表格样式设置

SPSS 为读者预设了一些格式的表格，每个格式的表格都有各自的特点，读者可以根据需要选择这些表格的样式。

##### 一般外观的特征设置

① 双击表格使其进入编辑状态。

② 按 **Format**→**TableLooks** 顺序单击菜单项，展开相应的对话框，见图 3-15。

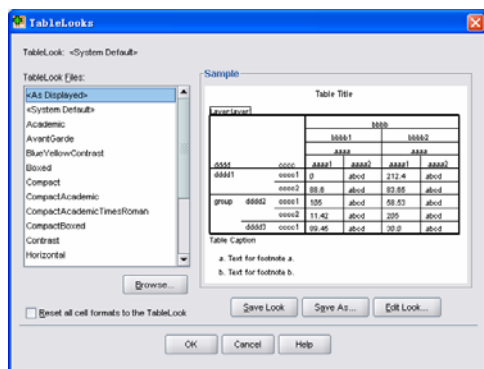


图 3-15 表格样式对话框

③ 在 **Table Look Files** 栏中选择表格样式文件，在 **Sample** 栏中观察所选定的表格样式文件代表的表格样式。

④ 如果想选择保存在文件中的其他表格样式，单击 **Browse** 对话框，选择文件。SPSS 表格样式文件是 **tlo** 格式，文件扩展名为 **tlo**。

⑤ 选择 **Reset all cell formats to the Table look**，将表格中所有编辑过的单元格重新设置成这里选择的表格样式。

⑥ 单击 **Save Look** 按钮打开另存为对话框。将当前选择的表格样式保存为当前选择的表格样式文件，以备需要时使用。

⑦ 单击 **Save As** 按钮打开另存为对话框，将当前选择的表格样式保存到指定的路径下的指定文件中。

⑧ 单击 **Edit Look** 按钮，显示 **Table Properties** 对话框。在该对话框中可以按需要对 **Table Looks** 对话框中选择的表格样式进行修改和编辑。

⑨ 单击 **OK** 按钮，所选择的表格变成在对话框中选择的样式。

#### 2. 表格样式的编辑

要对表格样式进行编辑，可以先使用 **TableLooks** 对话框选择一种基本样式，然后对所选择的样式进行修改。进入编辑样式对话框的途径有两个：

- 从 **TableLooks** 对话框中选择一种表格样式后，单击 **Edit Look** 按钮进入 **Table Properties** 对话框，见图 3-16。

- 按 **Format**→**Table Properties** 顺序单击菜单项，打开 **Table Properties** 对话框。

对话框中有 **General**（一般特性）、**Footnotes**（注脚）、**Cell Formats**（单元格格式）、**Borders**（边框）、**Printing**（打印）5 个功能选项卡，读者从这 5 个方面修饰表格。

## (1) 常规特性设置

① 在对话框 General 选项卡的 General 栏中选择 Hide empty rows and columns, 表格中的空行或空列将被隐藏起来。

② 在 Row Dimension Labels 栏内设置作为维度的变量, 例如方差分析中的因子变量的变量名和变量标签显示的位置。

- In Corner, 变量名显示在表格左上角单元格中。变量标签显示在变量名旁边, 一般显示在变量名右边, 如图 3-17(a)所示。

- Nested 选项, 以嵌套方式显示, 如图 3-17(b)所示。

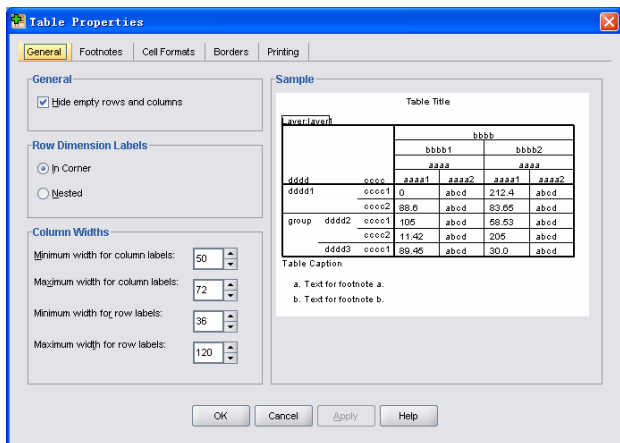


图 3-16 表格样式编辑对话框的一般特性设置

Dependent Variable	WUTERI	子宫重	
MOUSE 鼠	ETROGE	Mean	Std.
1 A	1 0.2	106.00	
	2 0.4	116.00	
	3 0.8	145.00	
Total		122.33	

Dependent Variable	WUTERI	子宫重	
MOUSE 鼠	1 A	ETROGEN	1 0.2
	2 0.4	雌激素	2 0.4
	3 0.8		3 0.8
	Total		
2 B	ETROGEN	雌激素	1 0.2

(a)

(b)

图 3-17 行维度标签显示位置

③ 在 Column Widths 栏中以像素点为单位设置列、行的极限宽度。

- Minimum width for column labels 框中设置最小列宽度。

- Maximum width for column labels 框中设置最大列宽度, 此值必须大于最小列宽度。

- Minimum width for row labels 框中设置最小行宽度, 此值必须小于最大行宽度。

- Maximum width for row labels 框中设置最大行宽度, 此值必须大于最小行宽度。设置完成后, 单击 Apply 按钮, 然后单击 OK 按钮, 对所选择的表格即刻产生效果。

## (2) 脚注设置

在输出的表格中常常需要添加脚注。在 Table Properties 对话框中的 Footnotes 选项卡中进行设置, 见图 3-18。

① Number Format 栏, 设置脚注方式。在 Sample 栏内看设置的实际效果。

- Alphabetic, 使用字母作为脚注标记。

- Numeric, 使用数字作为脚注标记。

② 在 Marker Position 栏, 设置脚注位置。

- Superscript, 脚注标记为上标, 显示在被标记对象的右上角。

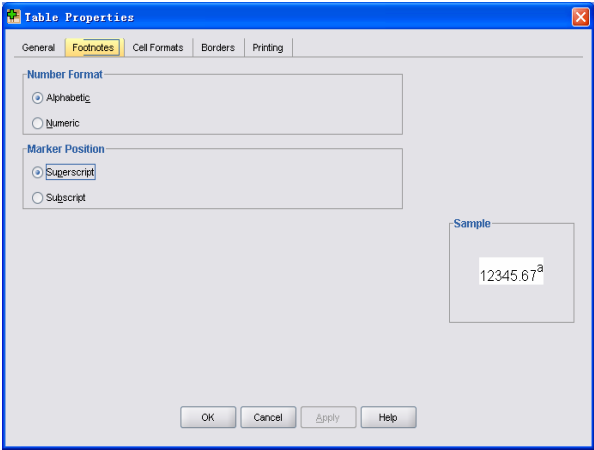


图 3-18 脚注标记选项卡

志)、Column Labels (列标志)、Data (数据)、Caption (表格下面注解标题) 和 Footnotes (表格下面注解文字, 即脚注)。

可以通过 Text 栏设置字体颜色的方法, 在 Sample 样例中看到各项代表的表格区域。

② 在 Text 栏设置字体、字号、加粗、倾斜、加下画线和改变字体颜色。

③ Alignment 栏, 设置表格中指定元素的对齐方式, 按钮自左至右分别为:

- 数字、日期右对齐, 所选择的其他元素在单元格中左对齐;
- 所选区域中的文字、数字对齐到单元格的左边界;
- 所选区域中的内容居中对齐;
- 所选区域中的文字、数字对齐到单元格的右边界;
- 所选区域中的小数点距右边界的距离为指定的距离;
- 指定单元格中数字中的小数点与右边界的距离, 单位是点、英寸或厘米, 这个单位在 Options 对话框的 General 选项卡中指定;

- 所选区域单元格中的内容对齐到上边界;
- 所选区域单元格中的内容垂直居中;
- 所选区域单元格中的内容对齐到下边界。

• Subscript, 脚注标记为下标, 显示在被标记对象的右下角。

单击 Apply 按钮, 再单击 OK 按钮, 设定的脚注即刻对所选择表格中的脚注生效。

(3) 单元格格式的设置

单击 Cell Format 选项卡, 在该选项卡中设置单元格格式, 见图 3-19。

① Area 选项框中选择要编辑哪个区域的单元格格式: Title (标题)、Layer (层)、Corner Labels (左上角单元格)、Row Labels (行标志)

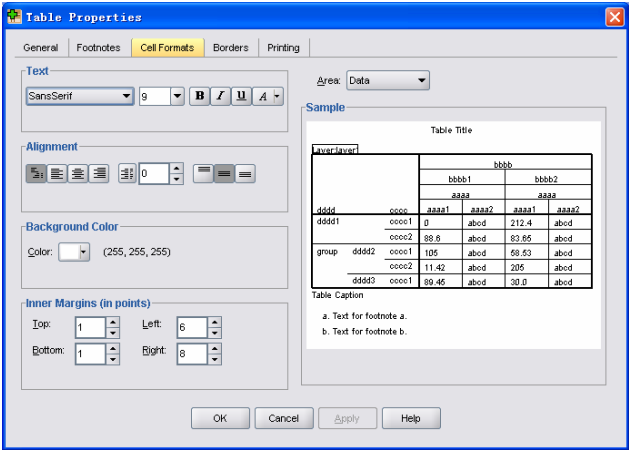


图 3-19 单元格格式选项卡

④ Background Color 框中单击向下箭头按钮, 显示调色板, 选择背景颜色。

⑤ 在 Inner Margents (in points) 栏中设置单元格内容与 Top (顶)、Bottom (底)、Left (左边界) 和 Right (右边界) 的距离。

#### (4) 设置边框格式

单击 Borders 选项卡, 如图 3-20 所示, 设置表格边框格式。表格边框指表格各位置上的表格线。

选项卡左边 Border 栏中列出的是各边框线的名称。在该栏中选择一种表格线。在 Border 栏下面左侧的下拉列表中选择线形, 在右边的下拉列表中选择线的颜色。在设置了不同线形或颜色后, 可以在右边的预览栏中看出 Border 栏中的选项指的是哪些边框线, 以及所设置的效果。

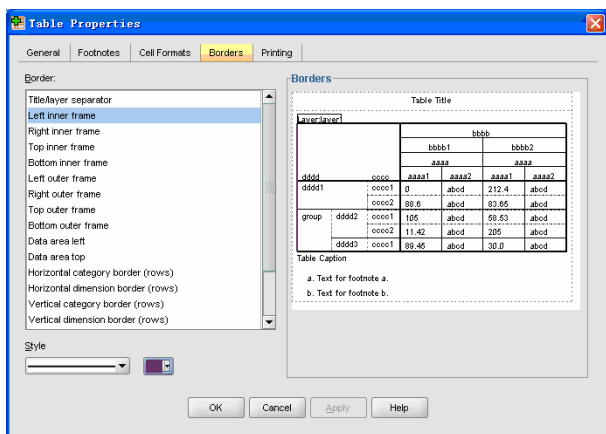


图 3-20 边框线格式选项卡

#### (5) 设置打印参数

单击 Printing 选项卡, 如图 3-21 所示, 设置有关打印的参数; 选项及其含义如下。

① Print all layers, 打印各层上的表格。选择此项, 激活 Print each layers on separate page 复选项, 如果选择该复选项, 每层表格打印在一页上。

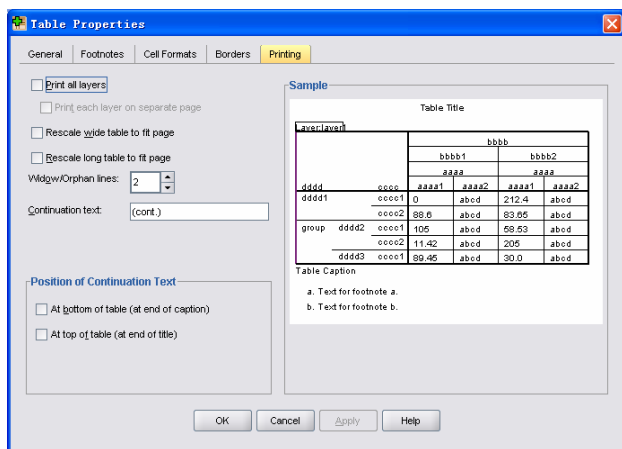


图 3-21 打印参数设置选项卡

⑤ 在 Continuation 后面的编辑区中输入一个标志性文字, 默认的是“Cont.”。当要打印的表格对设置的页来说太长, 需要打印在两页上时, 将 Continuation 后面编辑区中

② Rescale wide table to fit page, 压缩一个过宽的表格, 保持表格的纵横比, 以便在打印时适应在页面设置中设置的页宽。

③ Rescale long table to fit page, 压缩一个过长的表格, 保持表格的纵横比, 以便在打印时适应在页面设置中设置的页长。

④ Window/Orphan lines 参数框, 如果一个表格对所设置的页来说太长或太宽时, 该设置可以规定一个打印区中包含的最小行数和列数。

的文字打印在两页接续之处。

⑥ Position of Continuation Text 栏，设置在⑤中定义的接续文字显示（打印）的位置。表示接续的文字只在打印时出现，也可以使用打印预览观察到。

- At bottom of table (at end of caption)，表示接续的文字显示在表格底部。如果表格已经有了标题，接续文字加在标题后面。

- At top of table (at end of title)，表示接续的文字显示在表格顶部。如果表格已经有了标题，接续文字加在标题后面。

### 3.2.4 输出信息的复制与打印

如果撰写论文需要的分析结果数据在输出表格中，可将文字、表格复制到用 Word 撰写的论文中，可以使用选择、复制、粘贴的方法。但要注意以下两点：

(1) 直接将表格粘贴到 Word 文档中，其结果仍是 Word 表格，可以使用 Word 表格功能进行编辑和调整。

(2) 可以粘贴到 Windows 附件的画图窗口，然后剪切出来，再到 Word 窗口单击 Edit 菜单中的“选择性粘贴”命令，将剪贴板中的表格作为图片粘贴到 Word 文件中。可以节省所占的存储容量。粘贴后可以使用图片工具栏对表格图进行编辑。

如果选中表格后，单击右键菜单中的 Copy Objects 项，再粘贴到 Word 文档中。这个表格也是图片格式，可以使用 Word 中的图片工具栏中的各种工具对表格进行编辑。

**注意：**如果表格太宽，可以在复制之前先调整表格宽度，或隐藏不必要的数据列。

打印的参数设置与操作可以参考 Windows 的打印设置与操作。

## 习 题 3

1. 导航器的作用是什么？
2. 输出表格的数据能任意改变吗？为什么？
3. 怎样组织输出内容？
4. 表格太长，一页的宽度不能显示全部内容怎么办？

## 第 4 章 随机变量与分布函数的应用

### 4.1 随机变量与分布函数

#### 4.1.1 随机变量及其概率分布

##### 1. 随机变量

表示随机事件取值的变量就称为随机变量。

要对随机变量进行分析往往要对随机变量的取值数量化，即每个随机事件都使用数字来表示。例如，掷骰子时，用点数表示 6 个随机事件；合格产品用 1 表示，不合格产品用 0 表示；对某产品喜欢用 1 表示，不喜欢用 0 表示；喜欢程度：很喜欢、喜欢、无所谓、不喜欢、很不喜欢分别用 1、2、3、4、5 表示。

随机变量根据其取值的类型分为离散型随机变量和连续型随机变量。

随机变量是取有穷多值或可列无穷多值的称为离散型随机变量。例如对某品牌牙膏喜欢的程度很喜欢、喜欢、无所谓、不喜欢、很不喜欢分别用 1、2、3、4、5 表示。

可以取某区间中或某些区间中任何值的随机变量称为连续型随机变量。例如 1000 个成人的样本中的身高可以取 1.4~2.0m 之间任何值、体重可能取值在 30~100kg 等。

##### 2. 离散随机变量的概率分布

离散型随机变量的取值是有限的或可列无限的，如果知道每个可能的取值的概率，就可以用表格、图形（如表示相对频数的柱型图）或公式、表格表达概率分布的状况。

离散型随机变量的概率分布表达为：

设  $x$  所有可能的不同取值为  $x_i \quad i=1, 2, \dots, n$ ，或可列无限的  $i=1, 2, \dots$

$$P(x_i) = p_i, \quad i=1, 2, \dots, n \quad \text{对可列无限的 } i=1, 2, \dots$$

离散型随机变量的重要性质是

$$\sum_{i=1}^n p_i = 1$$

常见的离散随机变量的概率分布有以下三种。

##### (1) 两点分布

两点分布又称伯努利分布，是二项分布的特例。重复实验只有两种互斥的事件，事件的发生与不发生。这两种事件的分布服从两点分布。也就是说，随机变量只能取两个

值, 事件发生取值 1, 不发生取值 0。例如掷骰子的正面与反面; 市场调查中对一件商品的态度: 购买与不购买等。两点分布的概率分布函数表示为

$$P(X=x)=\begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$$

与之有关的函数为:

PDF.BERNOULLI(*quant*, *prob*) 数值型函数, 函数值等于分布参数为 *prob* 的伯努利分布在 *quant* 处的概率值。

CDF.BERNOULLI(*quant*, *prob*) 数值型函数, 给出符合概率为 *prob* 的二项分布的随机变量值小于或等于 *quant* 的累积概率值。

RV.BERNOULLI(*prob*) 数值型函数, 函数值为一个来自伯努利分布具有指定概率参数 *Prob* 的随机数。

## (2) 二项分布

满足下列条件的分布为二项分布:

- ① 从总体中抽取  $n$  个单元组成样本 (即重复  $n$  次实验),
- ② 各次实验相互独立, 每次实验只能有两种互斥的结果, 某事件 A 发生与不发生。
- ③ 每次实验, 事件 A 发生的概率为  $\pi$ , 记做  $P(A)=\pi$ , 不发生的概率为  $1-\pi$ 。在  $n$  次实验中事件 A 发生  $m$  次的概率的分布为二项分布, 如果用  $x$  表示事件 A 发生次数的随机变量, 该概率的表达式为:

$$P(x)=C_n^x \pi^x (1-\pi)^{n-x} \quad k=1,2,3,\cdots,n$$

SPSS 的二项分布概率函数为 PDF.BINOM 函数。

- PDF.BINOM(*quant*, *n*, *prob*) 数值型函数, 每次实验成功的概率是 *prob* 时, 函数值为  $n$  次实验中的成功次数等于 *quant* 的概率。

- CDF.BINOM(*quant*, *n*, *prob*) 数值型函数, 当每次实验成功的概率是 *prob* 时, 函数值是一个  $n$  次实验中成功次数小于或等于 *quant* 的二项分布累积概率值。

- RV.BINOM(*n*, *prob*) 数值型函数, 函数值是一个来自具有指定试验次数  $n$  和概率参数 *prob* 的二项式分布的随机数。

二项分布要求总体率 (或样本率) 不能太小, 不能接近 0, 例如  $<0.01$ 。如果事件的发生需要很大的样本量, 即  $n$  很大, 一次发生的概率很小, 二项分布就趋近泊松分布了。

## (3) 泊松分布

如果某稀有事件发生次数用随机变量  $x$  表示,  $x$  的取值范围是  $k=0, 1, 2, 3, \cdots$  而且随机变量  $x=k$  的概率是

$$p(x=k)=\frac{\lambda^k}{k!} e^{-\lambda} \quad k=0, 1, 2, 3, \Lambda \quad \lambda > 0$$

则称随机变量  $x$  服从参数为  $\lambda$  的泊松分布。

如果  $x$  的平均值为  $\mu$ , 上式可以用下面公式表示泊松概率分布

$$p(x) = \frac{\mu^x e^{-\mu}}{x!} \quad k = 0, 1, 2, 3, \dots, \infty$$

其中  $e$  是欧拉常数, 是自然对数的底,  $e=2.71828\cdots$

与泊松分布有关的函数:

- **PDF.POISSON(*quant, mean*)** 数值型函数, 函数值是具有指定均值和概率参数的泊松分布, 值等于 *quant* 的概率。

- **CDF.POISSON(*quant, mean*)** 数值型函数, 函数值是具有指定的均值或概率参数的泊松分布, 随机变量值小于或等于 *quant* 的累积概率。

- **RV.POISSON (*mean*)** 数值型函数, 函数值是一个具有指定均值 *mean* 或 *rate* 参数的泊松分布的随机数。

### 3. 连续型随机变量的概率分布

连续型随机变量是在某个定义的区间内可以取任意实数的变量。度量这些量的单位在理论上是可以无限再分的。

连续型随机变量取任何值的概率都是 0, 只有在某个区间中的概率才可能不是 0。所以不能象离散型随机变量那样列出每一个值的相应概率。对连续型随机变量, 我们用密度函数形式来描述。连续型概率密度函数  $f(x)$  满足下列两个条件

$$f(x) \geq 0 \quad ①$$

$$\int_{-\infty}^{+\infty} f(x) dx = 1 \quad ②$$

与离散型的概率不同的是,  $f(x)$  不是概率, 而是概率密度函数。累计分布函数是在连续分布的随机变量  $X$  小于某值  $x$  的概率  $P$ , 即  $P(X \leq x)$  是以概率密度函数曲线在该区间的面积表示即。

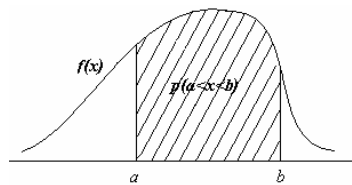


图 4-1 概率密度函数与概率

$$F(x) = \int_{-\infty}^x f(x) dx$$

当  $x=a$  时, 有概率  $F(a) = \int_{-\infty}^a f(x) dx$

当  $x=b$  时, 有概率  $F(b) = \int_{-\infty}^b f(x) dx$

随机变量在某区间上的概率是上述的公式②在某一个区间的积分。表示  $x$  值落在这个区间中的概率。

例如连续随机变量  $x$  落在  $(a,b)$  区间中的概率 (见图 4-1) 是

$$P(a < x < b) = \int_a^b f(x) dx$$



那么, 就有  $P(a < x < b) = F(b) - F(a)$

#### 4. 连续型随机变量的均值与标准差

连续型随机变量的均值定义为:

$$\mu = \int_{-\infty}^{+\infty} xf(x)dx$$

连续型随机变量的标准差定义为:

$$\sigma = \sqrt{\int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx}$$

连续型随机变量的常用概率分布有:

(1) 指数分布的概率密度函数为:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

与指数分布有关的 SPSS 函数为:

- PDF.EXP(*quant*, *shape*) 数值型函数, 函数值为形状参数为 *shape* 的指数分布在 *quant* 处的概率密度。

- CDF.EXP(*quant*, *shape*) 数值型函数, 函数值是具有给定的形状参数 *shape* 的指数分布的随机变量的值小于 *quant* 的累积概率。

- RV.EXP(*shape*) 数值型函数, 函数值是一个来自具有指定形状参数的指数分布的随机数。

(2) 正态分布的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < +\infty, \quad \sigma > 0$$

$\mu$  为随机变量  $x$  的均值,  $\sigma$  为标准差, 均为常数。随机变量服从均值为  $\mu$  标准差为  $\sigma$  的正态分布记做  $x \sim N(\mu, \sigma)$ ,

当均值 0, 标准差为 1 时的正态分布为标准正态分布。记做  $z \sim N(0,1)$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

可以通过  $z$  变换实现随机变量的标准化:  $z = \frac{x - \mu}{\sigma}$

与正态分布函数有关的 SPSS 函数为:

- PDF.NORMAL(*quant*, *mean*, *stddev*) 数值型函数, 函数值是具有指定的均值和标准差的正态分布, 在 *quant* 处的概率密度。

- CDF.NORMAL (quant,mean,stddev) 数值型函数，返回一个均值为 *mean*，标准差为 *stddev* 的正态分布的随机变量小于 *quant* 的累积概率。
- RV.NORMAL (mean, stddev) 数值型函数，函数值是一个具有指定均值 *mean* 和标准差 *stddev* 的正态分布随机数。

4.1.2 随机变量的函数

SPSS 中的随机变量的函数包括 7 类。概述于表 4-1 中。

随机变量和分布函数的关键字分前缀、后缀，前、后缀之间用圆点分隔。前缀指定分布的函数的归类，后缀指定分布。

随机变量和分布函数的自变量可以是常量也可以是变量。

如果要求函数自变量，对累积分布函数和概率密度函数和反分布函数的概率 *p*，必须出现在第一个，用 *x* 表示（*quant* 必须落在分布的合法值范围内）。

对随机变量和分布函数，必须指定分布参数作为对分布的说明。所有自变量都是实数。

表 4-1 7 类随机变量函数概述

类	解 释	数目
CDF	累积分布函数 CDF.d_spec(x,a,...)其值是累计概率 <i>p</i> ，具有指定的(d_spec)分布的连续随机变量落在 <i>x</i> 以下的累积概率；或对离散随机变量来说是在 <i>x</i> 处或 <i>x</i> 以下的概率	26
IDF	逆分布函数对离散分布不能用。 反分布函数 IDF.d_spec(p,a,...)函数值是 CDF.d_spec(x,a,...)= <i>p</i> 的具有(d_spec) 指定分布的 <i>x</i> 值	18
PDF	概率密度函数 PDF.d_spec(x,a,...)其值对连续随机变量来说是指定分布在 <i>x</i> 处的概率密度，值，对离散变量来说是具有指定分布的随机变量等于 <i>x</i> 的概率	23
RV	随机数发生函数 RV.d_spec(a,...)产生独立的具有指定分布的(d_spec)的观测量	22
NCDF	非中心累积分布函数 NCDF.d_spec(x,a,b,...)的值是一个具有指定的非中心分布的变量落在 <i>x</i> 以下的概率 <i>p</i> 。只对贝塔 (β) 分布、卡方分布、F 分布和学生化 T 分布可用	4
NPDF	非中心概率密度函数 NCDF.d_spec(x,a,...)的值是具有指定分布 (d_spec)的随机变量在 <i>x</i> 处的概率密度。只对贝塔 (β) 分布、卡方分布、F 分布和学生化 T 分布可用	4
SIG	尾概率分布函数（显著性函数）SIG.d_spec(x,a,...)的值是具有指定分布(d_spec)的变量大于 <i>x</i> 的概率 <i>p</i> 。它等于 1 减去累积分布函数值	2

1. 随机数函数（Random Numbers） 22 个

下面的函数根据指定的分布给出一个随机变量值。自变量是分布参数。

如果在数据文件中建立新变量时使用这些函数，变量值的个数等于数据文件中合法

的观测量数。

**注意：**函数名中的圆点是半角的圆点。

也可以事先在 Random Number Generators 随机数发生器对话框中通过设置一个种子，再由循环结构程序产生一系列符合一定分布的伪随机数。按 Transform→Random Number Generators 顺序，展开如图 4-2 所示对话框。在对话框中有以下两栏选项。

① Active Generator 有两个随机数发生器供选择。选择 Set Active Generator 用户设置工作发生器：

- SPSS 12 Compatible 与 SPSS12 兼容的发生器。如果需要再生成利用 12 版本或 12 版本以前发生器，基于一个种子值的随机化结果，就选择这个发生器。

- Mersenne Twister 一个新的更可靠的随机数发生器。

② Active Generator Initialization 现行发生器初始值。选择 Set Starting Point 可以由读者设置随机数发生器的初始种子值。有两个选项：

- Random，即由系统给出随机数作为产生随机数的种子。系统默认。

- Fixed Value 固定值，由读者设定。

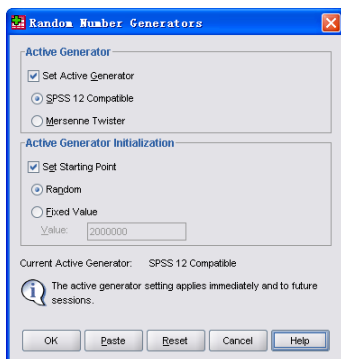


图 4-2 设置随机数发生器

随机变量函数如下：

(1) RV.BERNOULLI ( $prob$ ) 数值型函数，函数值是一个来自伯努利分布具有指定概率参数  $prob$  的随机数。

(2) RV.BETA ( $shape1, shape2$ ) 数值型函数，函数值是一个来自具有指定形状参数的 Beta 分布的随机数。

(3) RV.BINOM ( $n, prob$ ) 数值型函数，函数值是一个来自具有指定试验次数  $n$  和概率参数  $prob$  的二项式分布的随机数。

(4) RV.CAUCHY ( $loc, scale$ ) 数值型函数，函数值是一个来自具有指定位置  $loc$  和尺度  $scale$  参数的柯西分布的随机数。

(5) RV.CHISQ ( $df$ ) 数值型函数，函数值是一个来自具有指定自由度  $df$  的卡方分布的随机数。

(6) RV.EXP ( $shape$ ) 数值型函数，函数值是一个来自具有指定形状参数的指数分布的随机数。

(7) RV.F ( $df1, df2$ ) 数值型函数，函数值是一个来自具有指定自由度  $df1$ 、 $df2$  的 F 分布的随机数。

(8) RV.GAMMA ( $shape, scale$ ) 数值型函数，函数值是一个来自具有指定形状  $shape$  和尺度  $scale$  参数的伽马分布的随机数。

(9) RV.GEOM ( $prob$ ) 数值型函数，函数值是一个来自具有指定概率参数  $prob$  的几何分布的随机数。

(10) RV.HALFNM(*mean*, *stddev*) 数值型函数, 函数值是一个具有指定均值 *mean*、标准差 *stddev* 的半正态分布的随机数。

(11) RV.HYPER(*total*, *sample*, *hits*) 数值型函数, 函数值是一个来自具有指定参数的超几何分布的随机数。

(12) RV.IGAUSS(*loc*, *scale*) 数值型函数, 函数值是一个来自具有指定的位置参数 *loc* 和尺度参数 *scale* 的逆高斯分布的随机数。

(13) RV.LAPLACE(*mean*, *scale*) 数值型函数, 函数值是一个来自具有指定的均数 *mean* 和尺度 *scale* 参数的拉普拉斯分布的随机数。

(14) RV.LNORMAL(*a*, *b*) 数值型函数, 函数值是一个来自具有指定参数 *a*, *b* 的对数正态分布随机数。

(15) RV.LOGISTIC(*mean*, *scale*) 数值型函数, 函数值是一个来自具有指定的均数 *mean* 和尺度 *scale* 参数的逻辑斯蒂分布随机数。

(16) RV.NEGBIN(*threshold*, *prob*) 数值型函数, 函数值是一个具有指定阈值 *threshold* 和概率 *prob* 参数的负二项分布随机数。

(17) RV.NORMAL(*mean*, *stddev*) 数值型函数, 函数值是一个具有指定均值 *mean* 和标准差 *stddev* 的正态分布随机数。

(18) RV.PARETO(*threshold*, *shape*) 数值型函数, 函数值是一个具有指定阈值 *threshold* 和形状 *shape* 参数的帕雷托分布随机数。

(19) RV.POISSON(*mean*) 数值型函数, 函数值是一个具有指定均值 *mean* 或 *rate* 参数的泊松分布的随机数。

(20) RV.T(*df*) 数值型函数, 函数值是一个来自具有指定自由度 *df* 的学生 T 分布的随机数。

(21) RV.UNIFORM(*min*, *max*) 数值型函数, 函数值是一个属于具有指定最大值 *max* 和最小值 *min* 的均匀一致分布的随机数。另请参考 UNIFORM 函数。

(22) RV.WEIBULL(*a*, *b*) 数值型函数, 函数值是一个属于具有指定参数 *a*, *b* 的威布尔分布的随机数。

## 2. 概率密度函数 (PDF) 23 个

下列函数给出具有指定的分布, 在第一个自变量 *quant* 值处的密度函数的值。后面的自变量是分布参数。

**注意:** 每个函数名中的句点是英文半角的。

(1) PDF.BERNOULLI(*quant*, *prob*) 数值型函数, 函数值等于分布参数为 *prob* 的伯努利分布在 *quant* 处的概率值。

(2) PDF.BETA(*quant*, *shape1*, *shape2*) 数值型函数, 函数值等于形状参数为 *shape1*、*shape2* 的 beta 分布, 在 *quant* 处的概率密度值。

(3) PDF.BINOM(*quant*, *n*, *prob*) 数值型函数, 每次实验成功的概率是 *prob* 时, 函数

值为  $n$  次实验中的成功次数等于  $quant$  的概率。当  $n=1$  时, 该函数与 PDF.BERNOULLI 相同。

(4) PDF.BVNOR( $quant1, quant2, corr$ ) 数值型函数, 函数值为具有给定的相关系数  $corr$  的标准二元正态分布, 在  $quant1, quant2$  处的概率密度值。

(5) PDF.CAUCHY( $quant, loc, scale$ ) 数值型函数, 函数值为具有给定的位置参数  $loc$  和尺度参数  $scale$  的 Cauchy 分布在  $quant$  处的概率密度。

(6) PDF.CHISQ( $quant, df$ ) 数值型函数, 函数值为自由度为  $df$  的卡方分布在  $quant$  处的概率密度。

(7) PDF.EXP( $quant, shape$ ) 数值型函数, 函数值为形状参数为  $shape$  的指数分布在  $quant$  处的概率密度。

(8) PDF.F( $quant, df1, df2$ ) 数值型函数, 函数值为自由度为  $df1, df2$  的 F 分布在  $quant$  处的概率密度。

(9) PDF.GAMMA( $quant, shape, scale$ ) 数值型函数, 返回形状参数为  $shape$ , 尺度参数为  $scale$  的 Gamma 分布在  $quant$  处的概率密度。

(10) PDF.GEOM( $quant, prob$ ) 数值型函数, 返回概率值是当成功的概率是给定的  $prob$  时获得一次成功的实验数。

(11) PDF.HALFNRM( $quant, mean, stddev$ ) 数值型函数, 返回均值为  $mean$ , 标准差为  $stddev$  的半正态分布在  $quant$  处的概率密度。

(12) PDF.HYPER( $quant, total, sample, hits$ ) 数值型函数, 返回的数值是, 当从大小为  $total$  的,  $hits$  个对象具有指定特征的总体中随机选取样本  $sample$  时, 采样数中具有指定特征的对象数等于  $quant$  的概率。

(13) PDF.IGAUSS( $quant, loc, scale$ ) 数值型函数, 返回具有给定的位置参数  $loc$  和尺度参数  $scale$  的逆高斯分布, 在  $quant$  处的概率密度。

(14) PDF.LAPLACE( $quant, mean, scale$ ) 数值型函数, 返回具有指定均值  $mean$  和尺度参数  $scale$  的拉普拉斯分布, 在  $quant$  处的概率密度值。

(15) PDF.LNORMAL( $quant, a, b$ ) 数值型函数, 函数值是具有指定参数  $a, b$  的对数正态分布, 在  $quant$  处的概率密度值。

(16) PDF.LOGISTIC( $quant, mean, scale$ ) 数值型函数, 返回具有指定均值  $mean$  和尺度参数  $scale$  的 Logistic 分布, 在  $quant$  处的概率密度值。

(17) PDF.NEGBIN( $quant, thresh, prob$ ) 数值型函数, 函数值是当极限参数是给定的  $thresh$ , 成功的概率是  $prob$ , 获得一次成功的实验数等于  $quant$  的概率。

(18) PDF.NORMAL( $quant, mean, stddev$ ) 数值型函数, 函数值是具有指定的均值和标准差的正态分布, 在  $quant$  处的概率密度。

(19) PDF.PARETO( $quant, threshold, shape$ ) 数值型函数, 函数值是具有指定的阈值  $threshold$  和形状参数  $shape$  的帕累托分布, 在  $quant$  处的概率密度。

(20) PDF.POISSON(*quant*, *mean*) 数值型函数, 函数值是具有指定均值和概率参数的泊松分布, 值等于 *quant* 的概率。

(21) PDF.T(*quant*, *df*) 数值型函数, 函数值是具有指定的自由度 *df* 的学生化 T 分布, 在 *quant* 处的概率密度。

(22) PDF.UNIFORM(*quant*, *min*, *max*) 数值型函数, 函数值是具有指定的最小值 *min* 参数和最大值参数 *max* 的一致分布, 在 *quant* 处的概率密度。

(23) PDF.WEIBULL(*quant*, *a*, *b*) 数值型函数, 函数值是具有指定参数 *a*、*b* 的威布尔分布在 *quant* 处的概率密度。

### 3. 非中心分布的概率密度函数 (Noncentral PDF) 4 个

(1) NPDF.BETA(*quant*, *shape1*, *shape2*, *nc*) 数值型函数, 函数值是具有指定的形状参数 *shape1*, *shape2* 和非中心参数的 *nc* 的偏 Beta 分布, 在 *quant* 处的概率密度。

(2) NPDF.CHISQ(*quant*, *df*, *nc*) 数值型函数, 函数值是具有指定的自由度 *df* 和指定的非中心参数 *nc* 的偏卡方分布, 在 *quant* 的概率密度值。

(3) NPDF.F(*quant*, *df1*, *df2*, *nc*) 数值型函数, 函数值是具有指定的自由度 *df1*、*df2* 和非中心参数 *nc* 的偏 F 分布, 在 *quant* 处的概率密度。

(4) NPDF.T(*quant*, *df*, *nc*) 数值型函数, 函数值是具有指定的自由度 *df* 和非中心参数 *nc* 的偏学生化 T 分布, 在 *quant* 处的概率密度值。

### 4. 累积分布函数 (CDF) 26 个

下面的函数给出具有指定分布参数的随机变量值小于第一个自变量 *quan* 的累积概率, 分布类型由函数名决定, 后面的自变量是分布参数。

**注意:** 函数名中的圆点必须是英文半角圆点。

(1) CDF.BERNOULLI (*quant*, *prob*) 数值型函数, 给出符合概率为 *prob* 的二项分布的随机变量, 其值小于 *quant* 的累积概率值。

(2) CDF.BETA (*quant*, *shape1*, *shape2*) 数值型函数, 给出其值小于 *quant*, 来自具有给定的形状参数 *shape1*、*shape2* 的 Bate 分布的随机变量的累积概率值。

(3) CDF.BINOM (*quant*, *n*, *prob*) 数值型函数, 当每次实验成功的概率是 *prob* 时, 函数值是一个 *n* 次实验中成功次数小于或等于 *quant* 的二项分布累积概率值, 当 *n*=1 时, 该函数与 CDF.BERNOULLI 相同。

(4) CDF.BVNOR(*quant1*, *quant2*, *corr*) 数值型函数, 给出的函数值为来自二元标准正态分布的两个随机变量相关系数为 *corr*, 其值分别小于 *quant1*、*quant2* 的累计概率。

(5) CDF.CAUCHY (*quant*, *loc*, *scale*) 数值型函数, 函数值是具有给定的位置参数 *loc* 和尺度参数 *scale* 的柯西分布的随机变量, 其值小于 *quant* 的累积概率值。

(6) CDF.CHISQ (*quant*, *df*) 数值型函数, 函数值是具有给定的自由度 *df* 的卡方分布的随机变量, 其值小于 *quant* 累积概率。

(7) CDF.EXP (*quant*, *shape*) 数值型函数, 函数值是具有给定的形状参数 *shape* 的指

数分布的随机变量, 其值小于 *quant* 累积概率。

(8) CDF.F (*quant*, *df1*, *df2*) 数值型函数, 函数值是具有给定的自由度 *df1*、*df2* 的 F 分布的随机变量, 其值小于 *quant* 的累积概率值。

(9) CDF.GAMMA (*quant*, *shape*, *scale*) 数值型函数, 函数值是具有给定的形状参数 *shape* 和尺度参数 *scale* 的伽玛分布的随机变量, 其值小于 *quant* 的累积概率。

(10) CDF.GEOM (*quant*, *prob*) 数值型函数, 返回一个值小于 *quant* 的几何分布的累积概率; 即当成功概率为 *prob* 时, 获得一次成功的试验次数。

(11) CDF.HALFNM(*quant*, *mean*, *stddev*) 数值型函数。函数值是具有指定的均值 *mean*, 标准差 *stddev* 的半正态分布的随机变量, 其值小于 *quant* 的累积概率值。

(12) CDF.HYPER (*quant*, *total*, *sample*, *hits*) 数值型函数, 样品 *sample* 个事件是从大小为 *total* 的有 *hits* 个具有指定特性的总体中随机选择出来的情况下, 返回随机变量小于或等于 *quant* 的累积概率, 即具有指定特性的事件数。

(13) CDF.IGAUSS(*quant*, *loc*, *scale*) 数值型函数, 函数值为具有给定的位置参数 *loc* 和尺度参数 *scale* 的逆高斯分布的随机变量, 其值小于 *quant* 累积概率。

(14) CDF.LAPLACE (*quant*, *mean*, *scale*) 数值型函数, 返回来自均值为 *mean*, 尺度参数 *scale* 的拉普拉斯分布的, 随机变量值小于 *quant* 的累积概率。

(15) CDF.LNORMAL(*quant*, *a*, *b*) 数值型函数, 返回具有指定参数 *a*、*b* 的对数正态分布的, 随机变量值小于 *quant* 的累积概率值。

(16) CDF.LOGISTIC (*quant*, *mean*, *scale*) 数值型函数, 返回来自具有给定的均值 *mean* 和尺度参数 *scale* 的逻辑斯蒂分布的, 随机变量值小于 *quant* 的累积概率。

(17) CDF.NEGBIN (*quant*, *thresh*, *prob*) 数值型函数, 返回值小于 *quant* 的累积概率值, 即当阈值参数为 *thresh*, 成功的概率为 *prob*, 获得一次成功的实验次数。

(18) CDF.NORMAL (*quant*, *mean*, *stddev*) 数值型函数, 返回一个均值为 *mean*, 标准差为 *stddev* 的正态分布的, 随机变量值小于 *quant* 的累积概率。

(19) CDF.PARETO (*quant*, *threshold*, *shape*) 数值型函数, 返回阈值为 *threshold*, 形状参数 *shape* 帕雷托分布, 随机变量值小于 *quant* 的累积概率。

(20) CDF.POISSON (*quant*, *mean*) 数值型函数, 返回一个来自具有指定的均值或概率参数的泊松分布, 随机变量值小于 *quant* 的累积概率。

(21) CDF.SMOD(*quant*, *a*, *b*) 数值型函数, 返回属于学生化最大模, 具有指定参数 *a*、*b*, 随机变量值小于 *quant* 的累积概率。

(22) CDF.SRANGE(*quant*, *a*, *b*) 数值型函数。返回一个具有指定参数 *a*、*b* 的学生化值域分布, 随机变量值小于 *quant* 的累积概率值。

(23) CDF.T (*quant*, *df*) 数值型函数, 返回一个具有指定的自由度 *df* 的学生 T 分布的随机变量, 值小于 *quant* 的累积概率。

(24) CDF.UNIFORM (*quant*, *min*, *max*) 数值型函数, 返回一个具有指定的最小值 *min*

和最大值  $max$  参数的一致分布的随机变量, 值小于  $quant$  的累积概率。

(25) CDF.WEIBULL ( $quant, a, b$ ) 数值型函数, 返回一个具有指定的参数  $a$ 、 $b$  的威布尔分布的随机变量, 值小于  $quant$  的累积概率。

(26) CDFNORM( $zvalue$ ) 数值型函数, 返回一个具有均值为 0, 标准差为 1 的标准正态分布, 随机变量的值小于  $zvalue$  的概率。

#### 5. 非中心分布的累积概率密度函数 (Noncentral CDF) 4 个

(1) NCDF.BETA ( $quant, shape1, shape2, nc$ ) 数值型函数, 返回一个具有指定的形状参数  $shape1$ 、 $shape2$  和非中心参数  $nc$  的 Beta 分布的随机变量, 值小于  $quant$  的累积概率。

(2) NCDF.CHISQ ( $quant, df, nc$ ) 数值型函数, 返回一个具有指定的自由度  $df$ 、非中心性参数  $nc$  无偏卡方分布的随机变量, 值小于  $quant$  的累积概率。

(3) NCDF.F ( $quant, df1, df2, nc$ ) 数值型函数, 返回一个具有指定的自由度  $df1$ 、 $df2$  和非中心性参数  $nc$  的非中心 F 分布的随机变量, 值小于  $quant$  的累积概率。

(4) NCDF.T ( $quant, df, nc$ ) 数值型函数, 返回一个具有指定的自由度  $df$ 、非中心性参数  $nc$  非中心 T 分布的随机变量, 值小于  $quant$  的累积概率。

#### 6. 反分布函数 (Inverse DF) 18 个

下面的函数给出一个在指定的分布中的值, 这个分布的累积概率为第一个自变量  $prob$  的值, 其后的自变量是指定分布的参数。注意每个函数名由两部分组成, 圆点前是函数类名, 圆点后是分布名称, 括号内是自变量。当已知某分布累积概率值, 求随机变量值时使用此类函数。

(1) IDF.BETA ( $prob, shape1, shape2$ ) 数值型函数, 函数值为, 形状参数为  $shape1$ 、 $shape2$  的 Beta 分布的随机变量, 在累积概率为  $prob$  处的值。

(2) IDF.CAUCHY ( $prob, loc, scale$ ) 数值型函数, 函数值为位置参数  $loc$  和尺度参数  $scale$  的柯西分布的随机变量, 在累积概率为  $prob$  处的值。

(3) IDF.CHISQ ( $prob, df$ ) 数值型函数, 函数值为一个卡方值, 该卡方分布的自由度为  $df$ , 概率值为  $prob$ 。例如在 0.05 水平上 (累积概率为 95%), 自由度为 3 的卡方值为 IDF.CHISQ(0.95,3)。

(4) IDF.EXP ( $p, scale$ ) 数值型函数。函数值为按  $scale$  速度指数衰减的随机变量, 累积概率在  $p$  处的值。

(5) IDF.F ( $prob, df1, df2$ ) 数值型函数, 函数值为自由度为  $df1$ 、 $df2$  的 F 分布的随机变量, 累积概率  $prob$  的值。例如显著性概率在 0.05 水平上, 自由度分别为 3、100 的 F 值为 IDF.F (0.95,3,100)。

(6) IDF.GAMMA ( $prob, shape, scale$ ) 数值型函数, 函数值为形状参数为  $shape$  和尺度参数为  $scale$  的伽玛分布的随机变量, 累积概率为  $prob$  的值。

(7) IDF.HALFNRM( $prob, mean, stddev$ ) 数值型函数。函数值为一个具有指定的均值  $mean$  标准差  $stddev$  的半正态分布的随机变量, 累积概率为  $prob$  的值。



(8) IDF.IGAUSS(*prob*, *loc*, *scale*) 数值型函数。函数值为具有给出的位置参数 *loc* 和尺度参数 *scale* 的逆高斯分布随机变量, 累积概率是 *prob* 的值。

(9) IDF.LAPLACE(*prob*, *mean*, *scale*) 数值型函数, 函数值等于均值为 *mean* 和尺度参数为 *scale* 的拉普拉斯分布的随机变量, 累积概率为 *prob* 的值。

(10) IDF.LNORMAL(*prob*, *a*, *b*) 数值型函数, 函数值为有指定参数 *a*, *b* 的对数正态分布的随机变量, 累积概率为 *prob* 的值。

(11) IDF.LOGISTIC(*prob*, *mean*, *scale*) 数值型函数, 函数值等于均值为 *mean* 和尺度参数为 *scale* 的逻辑斯蒂分布的随机变量, 累积概率为 *prob* 的值。

(12) IDF.NORMAL(*prob*, *mean*, *stddev*) 数值型函数, 函数值为具有指定均值和标准差正态分布随机变量, 累积概率为 *prob* 的值。

(13) IDF.PARETO(*prob*, *threshold*, *shape*) 数值型函数, 函数值为阈值 *threshold*, 尺度参数 *scale* 的帕累托分布的随机变量, 在累积概率 *prob* 处的值。

(14) IDF.SMOD(*prob*, *a*, *b*) 数值型函数, 函数值是具有指定参数 *a*, *b* 的学生最大模数随机变量, 累积概率是 *prob* 的值。

(15) IDF.SRANGE(*prob*, *a*, *b*) 数值型函数, 函数值为具有指定参数 *a*, *b* 的学生化值域统计量, 累积概率是 *prob* 的值。

(16) IDF.T(*prob*, *df*) 数值型函数, 函数值为具有指定自由度 *df* 的学生 T 分布的随机变量, 其累积概率为 *prob* 的值。

(17) IDF.UNIFORM(*prob*, *min*, *max*) 数值型函数, 函数值为具有的最大值 *max*、最小值 *min* 的均匀分布的随机变量, 其累积概率为 *prob* 值。

(18) IDF.WEIBULL(*prob*, *a*, *b*) 数值型函数, 函数值是具有指定的参数的韦伯分布的随机变量, 其累积概率为 *prob*。

## 7. 尾概率分布函数 (Significance) 2 个

下列函数给出具有指定分布的随机变量大于第一个自变量 *quant* 的概率, 后边的自变量是分布参数。

(1) SIG.CHISQ(*quant*, *df*) 数值型函数, 函数值是具有 *df* 自由度的卡方分布, 大于自变量 *quant* 值的累积概率。

(2) SIG.F(*quant*, *df1*, *df2*) 数值型函数, 函数值是自由度为 *df1*、*df2* 的 F 分布的, 大于自变量 *quant* 值的累计概率。

## 4.2 随机变量与分布函数应用

### 4.2.1 符合分布要求的随机数的生成

【例 1】如要生成均值为 0, 标准差为 1 的正态分布的随机数 1000 个, 方法如下:

(1) 首先在数据编辑窗中输入序号变量 no，数值型变量，输入编号 1~1000。即设法制造 1000 个观测量。

(2) 顺序单击 Transform→Computer Variable 打开计算新变量对话框。

(3) 输入新变量名 RNorm，单击 Type & Label 打开定义标签对话框，见图 4-3 (a)。用中文填写标签。变量类型为数值型，选择 Numeric。单击 Continue 返回主对话框。

(4) 在 Compute Variable 对话框中的 Function group 栏中选择 Random Numbers 类。在 Functions and Special Variables 栏中选择 Rv.Normal 函数，单击向上箭头按钮，将该函数原型 RV.NORMAL(?,?)显示在 Numeric Expression 栏内。输入函数参数：均值 0、标准差 1 代替两个问号为 RV.NORMAL(0,1)，见图 4-3 (b)。

(5) 单击 OK 按钮，生成均值为 0、标准差为 1 的正态分布随机数，显示在数据窗中。见图 4-3 (c)。根据生成的正态随机数绘制直方图，绘图方法及结果见图 4-4。

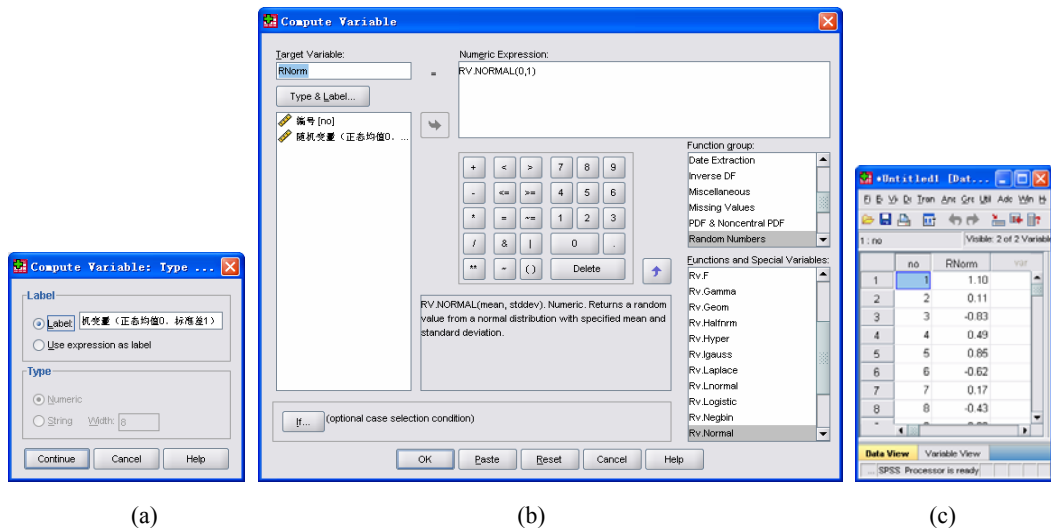


图 4-3 生成正态分布的随机数操作过程示意图

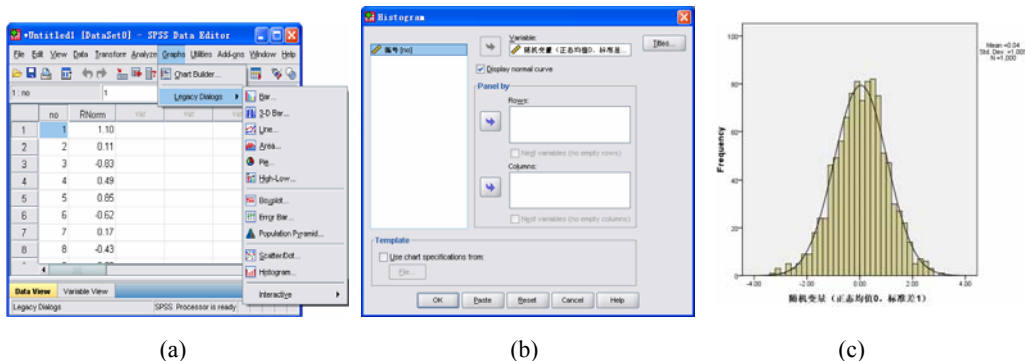


图 4-4 根据生成的正态随机变量的值绘图

(1) 按顺序单击 Graphs→Legacy Dialogs→Histogram 见图 4-4 (a)。展开绘制直方图的对话框见图 4-4 (b)。

(2) 选择随机变量,单击向右箭头按钮,将其移入 Variable 栏内,选择 Display Normal Curve。要求同时显示标准正态曲线。见图 4-4 (b)。

(3) 单击 OK 按钮,生成如图 4-4 (c)的直方图。实线为标准正态曲线。

## 4.2.2 概率密度函数与累积概率密度函数的应用

【例 2】离散随机变量及其分布的应用。

某体育专科学校的改革课题的调查表明,该类学校 75%的教师认为学生严重缺乏应该在中学阶段就掌握的基本技能。假定该校同意这一看法的总体率是  $\pi=0.75$ ,在某校抽取 20 名教师组成样本,问,20 人中有 11 人同意该意见的概率有多大?小于 10 人同意该意见的概率有多大?大于 10 人同意该意见的概率有多大?

解:同意与否是两个互斥事件,本例题中的实验结果数据属于二项分布。设同意该意见的人数为  $x$ 。原题为求  $P(x=11)$ 、 $P(x<10)$ 的累计概率和  $P(x>10)$ 的概率。

使用 SPSS 的 PDF.BINOM 函数解决这个问题  $P=0.75$ ,  $n$  次实验,  $n=20$ ,  $quant=11$ 。

① 建立一个变量  $x$ , 标签为“同意的人数”,输入数据 1~15。

② 建立另一个变量  $p$ , 标签为“同意的概率”输入数据为 15 个相同的数据 0.75。

见图 4-5 (a)。

③ 顺序单击 Transform→Compute Variable 打开计算变量对话框见图 4-5 (b)。

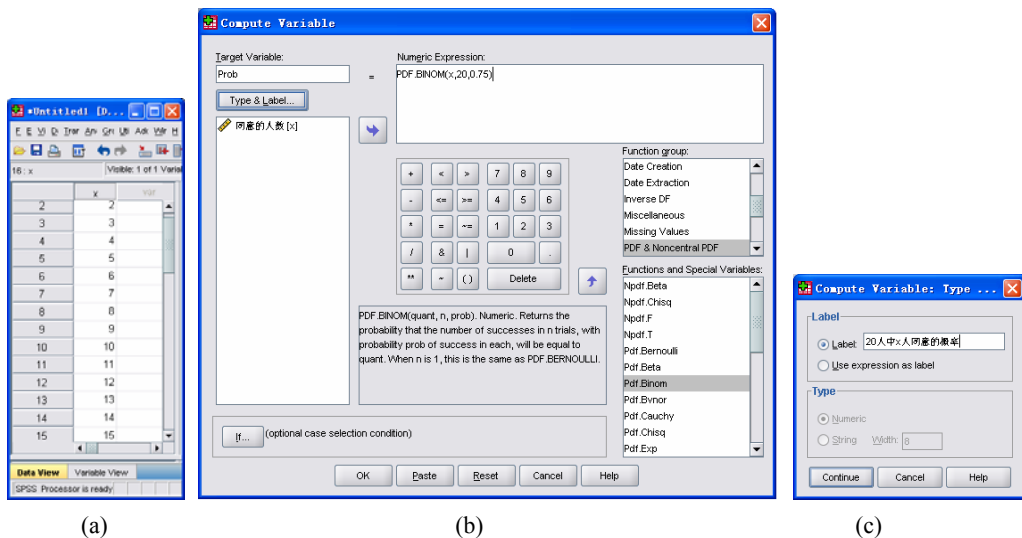
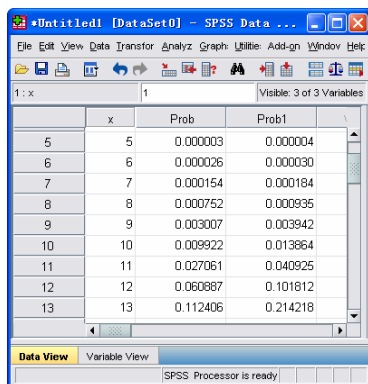


图 4-5 用 SPSS 函数求概率的方法

- Target Variable 栏内输入新变量名 prob。
  - 单击 Type & Label 按钮打开对话框见图 4-5
- (c) 设置变量类型为 Numeric 数值型；

- 在 Function Group 栏中选择 PDF & Noncentral PDF 类，在 Functions Special Variables 栏中选择 Pdf.Binom 函数，将函数送入 Numeric Expression 栏内。

- 在函数名后的括号中按顺序输入：“x,20,0.75”。函数显示为：PDF.BINOM(x,20,0.75)
- 单击 OK 按钮，在数据窗中生成变量 prob 的值。见图 4-6。



	x	Prob	Prob1
5	5	0.000003	0.000004
6	6	0.000026	0.000030
7	7	0.000154	0.000184
8	8	0.000752	0.000935
9	9	0.003007	0.003942
10	10	0.009922	0.013864
11	11	0.027061	0.040925
12	12	0.060887	0.101812
13	13	0.112406	0.214218

图 4-6 SPSS 计算结果

④ 在数据窗中查看  $x=11$  时的 prob 值为 0.027061，即 20 人中有 11 人同意的概率为 2.71%。（在变量观察窗中将变量的小数位数增加为 5 或 6）见 data04-01。

⑤ 0.00992，即 0.992% 是 20 人中 10 人同意的概率。

⑥ 与上述同样方法，建立新变量 Prob1 调用累积分布函数 CDF.BINOM(x,20,0.75)，得到新变量值，在数据窗中查看  $x=10$  的值为 0.013864，见图 4-6，数据文件见 data04-01。多于 10 人同意该观点的概率为：

$1-0.013864=0.986136$ 。即 98.61%。

【例 3】连续随机变量及其分布的应用。

在市场调查中，顾客对折扣优惠有不同看法和态度。一项调查对使用和不使用折扣优惠券的某种品牌的饮料价格进行比较。得出平均差价为 5.5 角，标准差为 3.5 角。假定差价  $x$ （角）服从正态分布。求差价大于 10 角的概率、大于 5 角的概率，以及小于 0 角的概率。注意，钱不是连续型的随机变量，这里只是介绍一种解决问题的思路与方法。

首先分析，大于 10 角的概率就是 1 减去差价小于等于 10 角的累积概率，同样，大于 5 角的概率就是 1 减去差价小于等于 5 角的累积概率。因此应该选用累积概率函数解决此问题。

(1) 建立变量  $x$ ，数值型，输入数据 0~10。

(2) 单击 Transform→Compute Variable 打开计算变量对话框。

① 在 Target Variable 栏输入变量名 Cp，单击 Type & Label 按钮，在打开的对话框中定义变量类型为数值型，变量标签为“累积概率”，见图 4-7(a)。

② 在 Function group 栏选择 CDF & Noncentral CDF 类，在 Functions and Special Variables 栏内选择 Cdf.Normal 函数，单击向上箭头按钮，将其移到 Numeric Expression 栏中。见图 4-7 (b)。

在数据编辑窗的变量观察窗口增加变量 Cp 的小数位数。见图 4-7 (c)。

- ③ 在函数自变量位置的三个问号处，分布输入  $x$ 、5.5、3.5。为  $Cdf.Normal(x,5.5,3.5)$ ，见图 4-7(b)。单击 OK 按钮。
  - ④ 在数据编辑窗的变量观察窗口增加变量  $C_p$  的小数位数。见图 4-7 (c)。
  - ⑤ 数据观察窗计算结果见图 4-8。数据见 data04-02。
- $x=10$  时， $C_p=0.90073$ ； $x=5$  时， $C_p=0.44320$ ； $x=0$  时， $C_p=0.05804$ ；因此差价大于 10 角（1 元）的概率为： $1-0.90073=9.927\%$ ；差价大于 5 角（1 元）的概率为： $1-0.44320=55.68\%$ ；差价小于 0 的概率为 5.804%。

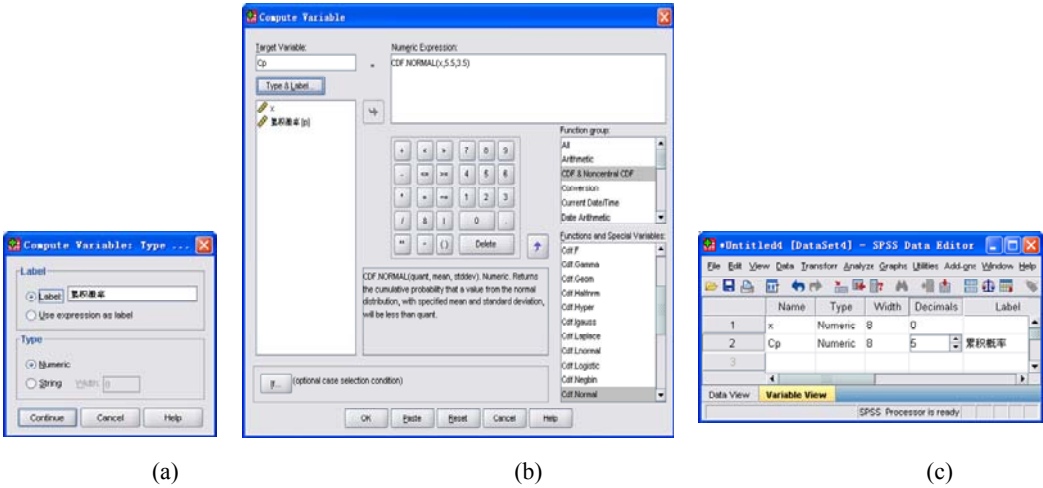


图 4-7 求累积概率的过程示意图

	x	Cp	var	var	var
1	0	0.05804			
2	1	0.09927			
3	2	0.15866			
4	3	0.23753			
5	4	0.33412			
6	5	0.44320			
7	6	0.55680			
8	7	0.66588			
9	8	0.76247			
10	9	0.84134			
11	10	0.90073			

图 4-8 计算结果

## 习 题 4

1. 某汽车公司的汽车销售量在过去 300 天的营业时间里, 有 55 天销售量为 0; 有 118 天销售量为 1; 有 70 天为 2 辆; 40 天为 3 辆; 10 天为 4 辆; 7 天为 5 辆。以过去 300 天的销售为历史数据, 问一天中售出 0、1、2、3、4、5 辆汽车的概率是多少? 以此验证离散型随机变量的概念与性质。

2. 用 RV.BERNOULLI 函数生成符合伯努利分布, 概率为 0.4 的 1000 个随机数; 用 RV.LNORMAL 函数, 生成参数  $a=0.2$ ,  $b=0.5$  的符合对数正态分布的随机数 1000 个, 并做直方图。

3. 某仪器上的部件长度要求非常严格, 要求在 0.304~0.322cm 之间, 某生产厂家生产的部件近似服从均值为 0.3015, 标准差为 0.0016 的正态分布。求该生产厂家不合格率是多少。

经改进, 该厂产品近似服从均值为 0.3146, 标准差为 0.0030, 不合格率为多少。

## 第 5 章 日期和时间函数及其运算

### 5.1 日期时间函数

#### 5.1.1 SPSS 日期时间概述

SPSS 的日期时间函数都借用固定数值进行转换，这个固定数值是 1582 年 10 月 14 日 0 时 0 分 0 秒。函数自变量无论 *timevalue* 还是 *datevalue* 都是以这个日期时间为基准的。如果以一个数值型变量作日期时间函数的自变量，日期时间函数将自变量的值看作自 1582 年 10 月 14 日 0 时 0 分 0 秒算起的秒数。在 SPSS 中输入 1582/10/14 以前的日期，系统自动将其转换为缺失值。因此如果把两个日期型变量直接运算时，要注意计算结果的类型和使用下面的函数进行转换后的数值所代表的含义。

#### 5.1.2 日期时间常量与变量

##### 1. 日期常量

日期常量的表示方法很多。SPSS 适应世界上不同国家、地区表示日期和时间的习惯，有多达 25 种表示方法。见表 5-1。示例中显示了我们可以直接使用的日期和时间的输入方法。中国人习惯的年、月、日顺序的表示方法，分为 2 位年和 4 位年两种。年、月、日之间用斜杠分隔。见表中灰色底纹的两行。

表 5-1 日期型常量格式及示例

格 式	说 明	示 例
dd-mmm-yyyy	日（2 位）-月份（英文）-年（4 位）	15-AUG-1945 23-DEC-2008
dd-mmm-yy	日（2 位）-月份（英文）-年（2 位）	15-AUG-45, 23-DEC-95
mm/dd/yyyy	月份（2 位）/日（2 位）/年（4 位）	08/15/1945, 12/23/1995
mm/dd/yy	月份（2 位）/日（2 位）/年（2 位）	08/15/45, 12/23/95
dd.mm.yy yy	日（2 位）.月份（英文）.年（4 位）	08.15.1945, 12.23.95
dd.mm.yy	日（2 位）.月份（英文）.年（2 位）	08.15. 45
yyyy/mm/dd	年（4 位）/月（2 位）/日（2 位）	2008/07/07

(续表)

格 式	说 明	示 例
yy/mm/dd	年（2 位）/月（2 位）/日（2 位）	08/08/15
yyddd	年（2 位）日数（自元月 1 日算起）	45227, 95
yyyyddd	年（4 位）日数（自元月 1 日算起）	1945227, 1995
q Q yyyy	季度 Q 年（4 位）	3Q1945, 4Q1995
q Q yy	季度 Q 年（2 位）	3Q45, 4Q95
mmm yyyy	月份（英文）年（4 位）	AUG1945, DEC1995
mmm yy	月份（英文）年（2 位）	AUG45, DEC95
ww WK yyyy	周数 “WK” 年（4 位）	33 WK 1945, 52 WK 1995
ww WK yy	周数 “WK” 年（2 位）	33 WK 45, 52 WK 95
Monday, Tuesday...	直接输入英文的星期几	Friday
Mon, Tue, Wed...	直接输入星期几的英文缩写	FRI
January, February...	直接输入英文月份	August, December
Jan, Feb, Mar...	直接输入英文月份缩写	AUG, DEC
dd-mmm-yyyy hh:mm	日（2 位）-月（英文月份缩写）-年（4 位） 时（2 位）:分（2 位）	11-AUG-1945 11:10
dd-mmm-yyyy hh:mm:ss	日（2 位）-月（英文月份缩写）-年（4 位） 时（2 位）:分（2 位）:秒（2 位）	11-AUG-1945 11:10:35
dd-mmm-yyyy hh:mm:ss.ss	日（2 位）-月（英文月份缩写）-年（4 位） 时（2 位）:分（2 位）:秒（2 位）.百分秒	11-AUG-1945 11:10:35.30
hh:mm	时（2 位）:分（2 位）	11:30, 08:50
hh:mm:ss	时（2 位）:分（2 位）:秒（2 位）	11:08:05, 08:15:25
hh:mm:ss.ss	时（2 位）:分（2 位）:秒（2 位）.百分秒	11:08:05.80, 08:15:25.45
ddd hh:mm	日数 时（2 位）:分（2 位）	128 08:50
ddd hh:mm:ss	日数 时（2 位）:分（2 位）:秒（2 位）	128 08:50:30
ddd hh:mm:ss.ss	日数 时（2 位）:分（2 位）:秒（2 位）.百分秒	128 08:50:30.78

注：“m”在年与日（字母 y 与 d）之间表示“月”份，在时与秒（字母 h 与 s）之间表示“分”。“mmm”三个字母表示要求书写英文月份单词的前三个字母组成的缩写。“ddd”三个字母 d 表示要求用从元月一日算起的日数表示日期。

指定了日期型变量的格式，不一定在输入时就使用指定的格式输入。可以输入用“/”或“-”作分隔符的具体日期，回车后，系统自动将输入的日期转换为指定格式，显示在单元格中。



2. 日期时间变量

日期时间变量的输出格式见表 5-2

表 5-2 日期时间型变量输入/输出格式

格式类型	说明	最小 w		最大 w	最大 d	一般格式	举例
		输入	输出				
DATEw	国际通用	9	9	40		dd-mmm-yy	28-OCT-90
		10	11			dd-mmm-yyyy	28-OCT-1990
ADATEw	美国	8	8	40		mm/dd/yy	10/28/90
		10	10			mm/dd/yyyy	10/28/1990
EDATEw	欧洲	8	8	40		dd.mm.yy	28.10.90
		10	10			dd.mm.yyyy	28.10.1990
JDATEw	朱利安	5	5	40		yyddd	90301
		7	7			yyyyddd	1990301
SDATEw	可排序的日期*	8	8	40		yy/mm/dd	90/10/28
		10	10			yyyy/mm/dd	1990/10/28
QYRw	季度和年	4	6	40		q Q yy	4 Q 90
		6	8			q Q yyyy	4 Q 1990
MOYRw	月和年	6	6	40		mmm yy	OCT 90
		8	8			mmm yyyy	OCT 1990
WKYRw	星期和年	6	8	40		ww WK yy	43 WK 90
		8	10			ww WK yyyy	43 WK 1990
WKDAYw	一周的天	2	2	40		周内天的英文名	SU
MONTHw	月	3	3	40		月的英文名	JAN
TIMEw	时间	5	5	40		hh:mm	01:02
TIMEw.d		10	10	40	16	hh:mm:ss.s	01:02:34.75
DTIMEw	天数和时间	1	1	40		dd hh:mm	20 08:03
DTIMEw.d		13	13	40	16	dd hh:mm:ss.s	20 08:03:00
DATETIMEw	日期和时间	17	17	40		dd-mmm-yyyy hh:mm	20-JUN-1990 08:03
DATETIMEw.d		22	22	40	16	dd-mmm-yyyy hh:mm:ss.s	20-JUN-1990

合法日期或日期时间变量值无论是以什么格式输入的，转换成另一种日期或日期时间格式，都能正常显示。因为机内值是不变的，改变的只是输出（显示）格式。Date11(即 dd/mmm/yyyy)和 Date9(dd/mmm/yy)才是标准格式。有些函数只对标准格式有效。

### 5.1.3 日期时间函数

#### 1. 当前日期时间函数 (Current Date/Time)

(1) **\$Date**, 字符型函数。其值为 9 位的 dd-mmm-yy 形式的当前日期。年数 2 位。格式是 A9。也就是说它是字符型, 要进行运算, 必须转换成数值格式或日期格式。

(2) **\$Date11**, 字符型函数。其值是 11 位的 dd-mmm-yyyy 的当前日期。年数占 4 位的。格式 A11。字符型, 要进行算术运算, 必须转换为数值格式, 或日期格式。

(3) **\$JDate**, 其值为数值型的当前日期, 是用从 1582 年 10 月 14 日 (罗马教皇格利高利定的第一天) 算起的年数表示的当前日期。格式是 F6.0。

(4) **\$Time**, 其值为当前日期和时间。**\$TIME** 给出的是从 1582 年 10 月 14 日 24:00:00 到转换命令执行之间的秒数。格式是 F20。你可以把它显示为一个用不同日期格式的数值的日期, 也可以把它用在日期和时间函数中。

#### 2. 日期的算术运算函数 (Date Arithmetic)

(1) **DATEDIFF**(*datetime2*, *datetime1*, "unit") 数值型函数。计算两个日期/时间值之间的差, 并按指定的日期/时间单位 *unit* 返回一个整数 (截去任何小数部分)。当 *datetime2* 和 *datetime1* 是日期或时间格式变量 (或者是表示有效的日期时间的数值), 而 "unit" 是下列字符串之一。用引号括起的: years、quarters、months、weeks、days、hours、minutes、seconds。(年、季度、月、周、天、小时、分、秒), 表示差值转换后的时间单位。

(2) **DATESUM**(*datetime*, *value*, "unit", "method") 数值型函数。计算 *datetime* 指定的日期或时间格式的变量 (或者表示合法日期/时间的数值) 与日期时间值 *Value* 之和, *unit* 是括在引号中下列字符串值之一: years、quarters、months、weeks、days、hours、minutes、seconds (年、季度、月、周、天、小时、分、秒), 表示值 *value* 的单位。*method* 是可选的, 可以是 "rollover" 或 "closest"。

"rollover" 滚动的方法把超出的天放到下一个月。

"closest" 最近法使用本月中最近的合法日期, 这是默认的方法。

返回的值是表示成秒数的日期/时间值。要显示成日期/时间就要给变量赋予适当的格式。可以用转换函数将数值变量转换成日期/时间格式。

#### 3. 时间生成函数 (Date Creation)

这是一组数值型函数。函数值是将日期的某年、月、日、季度、周的数字的有效组合转变成自 1582 年 10 月 14 日 0 点 0 分 0 秒起至指定日期的秒数。自变量必须是整数。其中的 *year* 必须是 4 位大于 1582 的表示年的整数, *Month* 是在 1~13 之间的月份。实际上有效值应该是 1~12, 如果输入数值为 13, 则按下一年的 1 月计算。*Quarter* 是在 1~4 之间的季度值, *weeknum* 是 1~52 之间的周数值; *daynum* 是在 1~366 之间的日数值。函数值是数值型, 要显示成日期, 只要在 **Variable View** 窗中将变量类型改变为日期型 (Date 型变量)。

(1) DATE.DMY(*day,month,year*) 数值型函数。返回与 *day*、*month* 和 *year* 相应的日期值。

(2) DATE.MDY(*month,day,year*) 数值型函数。返回与 *month*、*day* 和 *year* 相应的日期值。

(3) DATE.MOYR(*month,year*) 数值型函数。返回与 *month*、*year* 相应的日期值。

(4) DATE.QYR(*quarter,year*) 数值型函数。返回与 *quarter*、*year* 相应的日期值。

(5) DATE.WKYR(*weeknum,year*) 数值型函数。返回与 *weeknum*、*year* 相应的日期值。

(6) DATE.YRDAY(*year,daynum*) 数值型函数。返回与 *year*、*daynum* 相应的日期值。

#### 4. 日期提取函数 (Date Extraction)

① 这一组函数的自变量 *datevalue* (*timevalue*) 可以是:

- 数值或已经赋值的数值型变量或数值型表达式, 将自变量的值看作 1582 年 10 月 4 日 24:00:00 秒算起的天数 (秒数)。

- 日期 (时间) 型变量、日期时间型表达式或日期、时间值, 机内值是 1582 年 10 月 14 日 24:00:00 秒算起, 到达自变量指定日期 (时间) 的间隔中的天数 (秒数)。

② 自变量是 *timevalue* 的函数是数值型函数, 函数值为数值型常量。要把函数值显示成日期或时间, 应该赋予该函数值日期时间格式。

(1) XDATE.DATE(*datevalue*) 数值型函数。从数值返回表现为日期的日期部分, 要把结果显示成日期, 需要把变量赋予日期格式。

实验表明, 因变量定义成数值型, 返回的是数值, 从 1582……到自变量指定的日期之间的秒数; 如果因变量定义成日期型, 函数值仍然是原日期不变。

(2) XDATE.JDAY(*datevalue*) 数值型函数。函数值为一年中的天数 (1~366 之间的整数)。

(3) XDATE.MDAY(*datevalue*) 数值型函数。函数值是从 *datevalue* 提取出其但表日期中月份的第几天 (在 1~31 之间的整数)。

(4) XDATE.MONTH(*datevalue*) 数值型函数。函数值是从 *datevalue* 提取出的月份 (1~12 之间的整数)。

(5) XDATE.QUARTER(*datevalue*) 数值型函数。函数值是自变量所代表日期所在的一年中的季度 (1~4 之间的整数)。

(6) XDATE.TDAY(*timevalue*) 数值型函数。从表现为时间间隔的自变量数值返回整的天数。

(7) XDATE.TIME(*datetime*) 数值型函数。从一个表现为时间或日期时间的值返回时间部分。要把结果显示成时间, 要赋予结果变量一个时间格式。

(8) XDATE.WEEK(*datevalue*) 数值型函数。函数值是自变量表达的日期在该年的周数 (1~53 之间的整数)。

(9) XDATE.WKDAY(*datevalue*) 数值型函数。函数值为自变量 *datevalue* 表达的日期

所在周中的天数（在1周日~7周六之间的整数）。

(10) XDATE.YEAR(*datevalue*) 数值型函数。函数值是4位整数的年数。

(11) YRMODA(*year,month,day*) 数值型函数。根据自变量 *year*、*month*、*day* 返回从1582年10月14日起到自变量 *year*、*month*、*day* 表示的日期的天数。

#### 5. 时间间隔生成函数 (Time duration creation)

(1) TIME.DAYS(*days*) 数值型函数。函数值为与自变量 *days* 指定的天数相应的时间间隔。自变量必须是数值型。要将结果显示成时间，就要赋予结果变量时间格式。函数值是与自变量值相应的秒数。例如 TIME.DAYS(3)结果是259200。当赋予结果变量时间格式 hh:mm:ss 时，结果为 72:00:00。

(2) TIME.HMS(*hours*) 数值型函数。函数值为与时间间隔变量 *hours* 指定的小时数相应的秒数。*hours* 的值必须是整数；所有自变量必须处理成或者都是正值，或者都是负值。要把它显示成时间，要赋予结果变量时间格式。在函数列表中该函数名为 TIME.HMS(1)。例如 TIME.HMS(48)值为172800，赋予其时间格式 hh:mm 时，结果为 48:00。在保持结果变量为数值型时自变量可以是负值。其他格式将显示为缺失值。

(3) TIME.HMS(*hours,minutes*) 数值型函数。函数值是与时间间隔自变量 *hours*、*minute* 相应的秒数。*hours* 必须是整数；*minutes* 必须是小于60的整数。要函数值显示为时间，则应赋予结果变量时间格式。自变量可以都是负值或都是正值。对于都是负值的自变量，结果变量只能是数值型，其他格式将显示为缺失值。在函数列表中该函数名为 TIME.HMS(2)。例如 TIME.HMS(96,30)，结果为347400。当赋予结果变量时间格式 hh:mm 时，显示 96:30。

(4) TIME.HMS(*hours,minute,second*) 数值型函数。函数值是与时间间隔自变量 *hours*、*minute*、*second* 相应的秒数。*hours* 必须是整数；*minutes* 必须是小于60的整数；*seconds* 可以包括小数，但必须小于60。要函数值显示为时间，则应赋予结果变量时间格式。自变量可以都是负值或都是正值。对于都是负值的自变量，结果变量只能是数值型，其他格式将显示为缺失值。在函数列表中该函数名为 TIME.HMS(3)。

例如 TIME.HMS(96,30,20.50)，结果为347420.50。当赋予结果变量时间格式 hh:mm:ss 时，显示 96:30:20。

#### 6. 时间间隔提取函数 (Time duration Extraction)

(1) CTIME.DAYS (*timevalue*) 数值型函数，返回给定时间（被看做秒数）值折合的天数（自1582年10月14日算起的天数），包括分数的天数。自变量 *timevalue* 时间值必须是一个数值或是 SPSS 格式的时间表达式，如 TIME.xxx 函数的计算结果。例如 CTIME.DAYS (10800)值为0.125天，或者 CTIME.DAYS (time.hms(3))，结果相同。

(2) CTIME.HOURS (*timevalue*) 数值型函数，返回给定时间值折合的带有小数部分的小时数。自变量时间值必须是一个秒数值或 SPSS 格式的时间表达式，如 TIME.xxx 函数创建的时间值或用 TIME 输入格式读取的数值。如 CTIME.HOURS(172830)结果为48.008

显示值的近似程度取决于数值格式的小数位数。

(3) **CTIME.MINUTES** (*timevalue*) 数值型函数，返回给定时间值折合的带有小数部分的分钟数。自变量时间值必须是一个数值或 SPSS 格式的时间表达式，例如 **TIME.xxx** 函数创建的时间值或用 **TIME** 输入格式读取的数值。

(4) **CTIME.SECONDS** (*timevalue*) 数值型函数，返回给定时间值折合的带有小数部分的秒数。自变量时间值必须是一个数值或 SPSS 格式的时间表达式，例如 **TIME.xxx** 函数创建的时间值或用 **TIME** 输入格式读取的数值。

(5) **XDATE.HOUR** (*datevalue*) 数值型函数，函数值为与自变量 *datetime* 相应的小时数，一个 0~23 之间的整数。自变量是描述时间或日期时间值。自变量可以是一个数值、时间或日期时间变量或者处理成时间或日期时间值的表达式。

(6) **XDATE.MINUTE** (*datevalue*) 数值型函数，函数值为表现为时间或日期时间的值相应的 0~59 间的分钟数。自变量可以是一个数值、时间或日期时间变量或者可以是一个已经处理成时间或日期时间值的表达式。

(7) **XDATE.SECOND**(*datetime*)数值型函数，函数值为表现为时间或日期时间的值相应的 0~59 间的秒数。自变量可以是一个数值、时间或日期时间变量或者可以是一个已经处理成时间或日期时间值的表达式。

(8) **XDATE.TDAY**(*timevalue*)数值函数，函数值是与描述时间或日期时间的数值相应的整天数。自变量可以是一个数值，时间格式的变量或者处理成时间间隔的表达式。

### 7. 与日期时间有关的转换函数

**NUMBER**(*stringDate*,**DATE11**) 数值型函数，把内容为标准格式 (dd-mmm-yyyy) 日期的字符串转换成描述该日期的秒数。如果字符串不能使用标准格式读取，函数值是系统缺失值。

第一个自变量是字符型，自变量的值为与 **Date11** 格式相应的日期。

如果我们定义了字符串格式的自变量，输入了与 dd-mmm-yyyy 相应的日期，可以使用该函数将字符串变量转换为日期变量。

## 5.2 日期时间函数的应用

### 5.2.1 日期时间型变量的格式转换

日期时间型变量与数值型变量的转换。

因为日期时间型变量在机内就是从 1582 年 10 月 14 日 24:00:00 算起到达变量值代表的日期时间中的秒数。计算机内就是一个数值。只是在显示方式上有所不同。因此日期时间型变量与数值型变量的转换只是显示方式的转换。只要在数据窗的变量窗中改变变量的类型即可。

【例1】将 data05-01.sav 中的日期型变量 birthday 转换为数值型变量，方法如下：

(1) 图 5-1 (a) 是数据窗中的日期型变量 birthday 显示成 4 位年、2 位月、2 位日的日期格式。再将该变量数据复制到新变量 birthday1 中。单击窗口左下角的 Variable Viewer，将窗口切换到变量观察窗。如图 5-1 (b)所示。

(2) 单击变量观察窗 birthday1 的 Type 列的类型 Date，打开如图 5-1(c)所示的 Variable Type 定义变量类型的对话框。

(3) 在 Variable Type 对话框中选择 Numeric 项，图 5-1 (c)所示。单击 OK 按钮。在如图 5-1 (d)所示的窗口中显示的数值是从 1582 年 10 月 14 日 24:00:00(10 月 15 日 0:00:00)算起到 birthday 变量值指定日期的秒数值。日期型变量到数值型变量转换完毕。

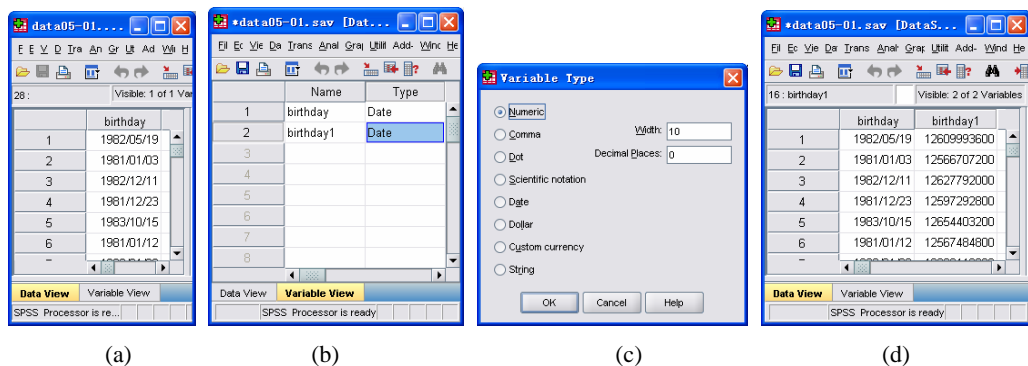


图 5-1 日期型变量转换成数值型变量的过程

转换后的数据文件是 data05-01a。

【例2】数据文件 data05-01 中的另一个变量 data1 是数值型变量。利用上述方法改变为日期型 Date 变量，并在 Variable Type 对话框中指定一种日期时间格式。见图 5-2(b)。

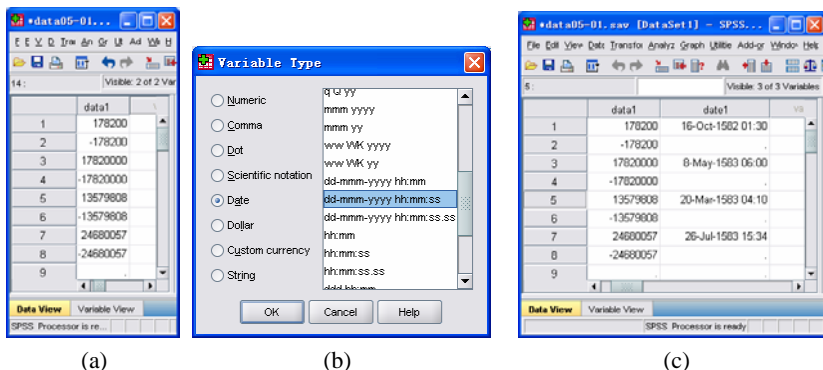


图 5-2 数值型变量转换成日期时间变量的过程

为比较，可以先将 data1 的数据拷贝到 date1 中。转换结果见图 5-2(c)。可以看出负数转换成日期型结果是缺失值。转换后的结果见数据文件 data05-02.sav。

【例 3】字符型变量与日期型变量的转换。

当前日期时间函数产生的是字符型常数。字符型是不能参与算术运算的。因此如果运算涉及由当前日期函数产生的变量时，必须先将显示成日期的字符型变量转换成日期型变量。图 5-3 (a)、(b) 是当前日期变量 currentdate 的数据观察窗口和变量观察窗口，该变量是字符型。见数据文件 data05-03.sav。

将显示为日期形式的字符型变量转换成数值型或日期型变量的方法可利用转换函数，例如使用 \$Date9 或 \$Date11 两个当前日期函数产生的 currentdate 就是

是字符型变量。转换为数值型变量的操作步骤是：

(1) 单击 Transform→Compute 打开 Compute Variable 对话框，见图 5-4。

(2) 在计算变量对话框中：

① Target Variable 栏中输入新变量名 currntdate1。

② 单击 Type & Label 按钮，打开如图 5-5 的定义新变量类型对话框。在 Label 栏选择 Label 项输入变量标签，可以使用中文。Type 栏内选择 Numeric，说明新变量是数值型。

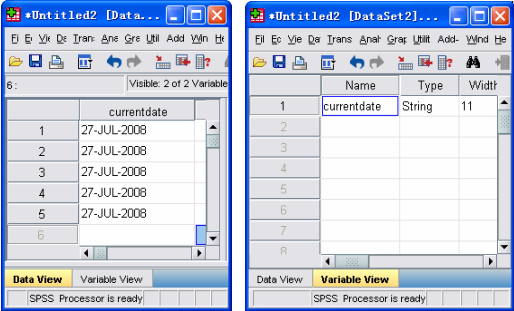
单击 Continue 按钮，返回 Compute Variable 对话框。

③ 在 Function group 栏中选择 Conversion，即转换类型的函数。

④ Functions and Special Variables 栏内选择 Number 函数，单击向上箭头按钮。显示为：Number(?,?)，光标停留在第一个问号处。

在变量列表中选择 currentdate 单击向右箭头按钮，在第二个问号处输入日期变量格式 Date9。就此形成等式：

Currentdate1=number(currntdate,data9)



(a) (b)

图 5-3 当前日期时间函数生成的字符型变量

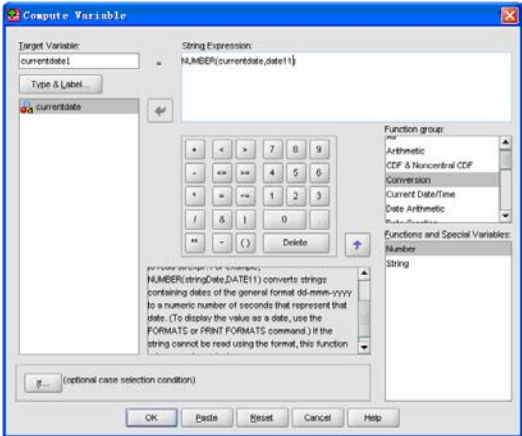


图 5-4 计算新变量对话框中转换变量类型

格式参数也可以是 date11 即：

```
Currentdate1=number(currntdate,date11)
```

⑤ 单击 OK 按钮，在数据窗中生成新变量 Currntdate1，转换后的数据见图 5-6。参见数据文件 data05-03a。

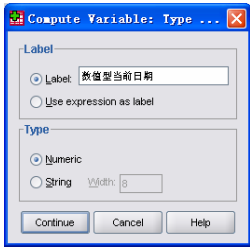


图 5-5 定义新变量类型图

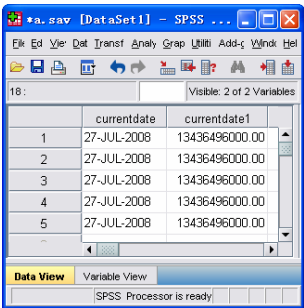
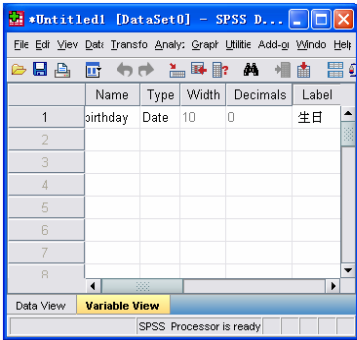


图 5-6 数值型新变量

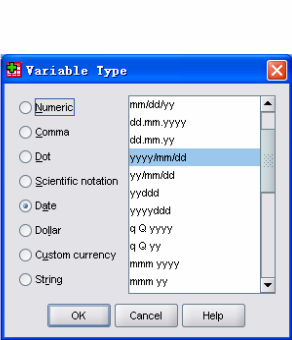
5.2.2 日期时间型变量的算术运算

【例 4】 业余体校某项运动的训练班花名册中记录了运动员的出生日期。计算到当前日期为止，这些队员的年龄。步骤如下：

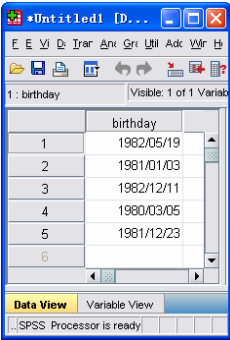
- (1) 在 Variable View 窗口建立一个变量名为 birthday 的变量。
- (2) 在 Type 栏内定义该变量为日期型。选择了下拉菜单中的 Date 打开选择格式对话框如图 5-7 (b)所示。在对话框中选择 yyyy/mm/dd。单击 OK 按钮返回 Variable View 窗口。自动显示变量宽度为 10。
- (3) 在 Data View 窗口中，按选择的 yyyy/mm/dd 格式输入运动员生日，如图 5-7 (c)所示。见 data05-04。



(a)



(b)



(c)

图 5-7 定义一个日期变量，输入日期数据



(4) 单击 Transform→Compute Variable 打开 Compute Variable 对话框; 在 Target Variable 栏内输入新变量名 Curda; 单击 Type & Label 按钮, 打开如图 5-8(a)所示的 Compute Variable:Type 对话框, 在 Label 栏输入变量标签“当前日期 0”; 在 Type 栏内选择 String。表明新变量是字符型。因为在该对话框内只能在数值型和字符型中选择一个, 而由于日期的表达中分隔符的存在, 不可能是数值型。宽度输入 10, 单击 Continue 按钮。

(5) 按上一节所述的方法将字符型的 Curda 当前日期变量转换为数值型。变量名为 Currdate。表达式为: NUMBER(curdat,date11)。见 data05-04a。

(6) 由于变量 birthday 是日期型变量, 而它的机内置是数值, Currdate 是数值型变量都是自 1582 年 10 月 14 日 24:00:00 算起的秒数, 所以可以进行算术运算。

(7) 调用日期计算函数, 得到年龄变量 Age。

单击 Transform→Compute Variable 打开相应的对话框, 见图 5-8(b)。

① 在 Target Variable 栏内输入新变量名 Age; 单击 Type & Label 按钮, 打开如图 5-8(a)所示的 Compute Variable:Type 对话框。在 Label 栏输入变量标签“年龄”; 在 Type 栏选择 Numeric; 单击 Continue 按钮返回 Compute Variable 对话框。

② 在 Compute Variable 对话框中, Function group 栏中选择 Date Arithmetic 类, 在 Functions and Special Variables 栏内选择计算日期差函数 Datediff, 单击向上箭头按钮, 该函数显示在 String Expression 栏中: Datediff(?,?,?)。

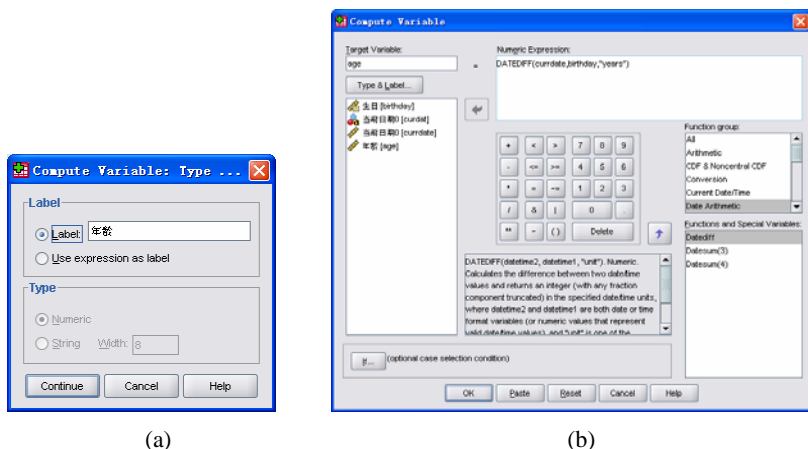


图 5-8 打开 Compute Variable 对话框, 生成当前日期变量

③ 在原始变量栏内选择当前日期 Currdate 变量作为第一个参数, 单击向右箭头按钮, 代替函数中第一个问号; 同样方法选择 birthday 作为第二个参数变量, 代替第二个问号; 在第三个问号处输入"years"作为第三个参数。注意必须带半角引号。在 Numeric Express 栏内显示调用函数: Datediff(Currdate,birthday,"years")。单击 OK 按钮。

④ 在数据观察窗中察看新变量 Age 的结果。如图 5-9 所示。

在输出窗中查看执行的语句是：

```
COMPUTE age=DATEDIFF(currdate,birthday,"years")
```

```
EXECUTE.
```

结果如图 5-9 所示。见数据文件 data05-04b。

【例 5】 班委会决定在每个月为在该月份过生日的同学时举办一次庆祝活动，班级花名册中记载着每个同学的生日，在数据窗中是变量 birthday。为了统计每个月有几个人过生日，需要把生日的月份提取出来。应该如何操作？

首先确定需要使用的是提取月份的函数 XDATE.MONTH(datevalue)，操作步骤如下：

(1) 单击 Transform→Compute Variable 打开

Compute Variable 对话框；

(2) 在 Target Variable 栏内输入新变量名 birthday1；单击 Type & Label 按钮，打开如图 5-8 (a) 的 Compute Variable :Type 对话框，在 Label 栏输入变量标签“生日月份”；在 Type 栏选择 Numeric；单击 Continue 按钮返回 Compute Variable 对话框。

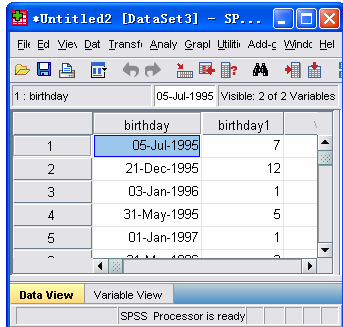


图 5-10 提取月份的结果

The screenshot shows the SPSS Data Editor window with the following data:

	birthday	currdate	currdate	age
1	1982/05/19 20-NOV-2009	13478054400.00	27.00	
2	1981/01/03 20-NOV-2009	13478054400.00	26.00	
3	1982/12/11 20-NOV-2009	13478054400.00	28.00	
4	1981/12/23 20-NOV-2009	13478054400.00	27.00	
5	1983/10/15 20-NOV-2009	13478054400.00	26.00	
6	1981/01/12 20-NOV-2009	13478054400.00	28.00	

At the bottom of the window, the status bar indicates "SPSS: Processor is ready".

图 5-9 带有年龄计算结果的数据窗

(3) 在 Function group 栏中选择 Date Extraction 类，在 Function and Special Variables 栏内选择提取月份的函数 XDATE.MONTH，单击向上箭头按钮，该函数显示在 String Expression 栏中：XDATE.MONTH(?)。

(4) 在原始变量栏内选择 birthday 作为函数自变量，单击向右箭头按钮，代替函数中的问号；显示为：XDATE.MONTH(birthday)。

(5) 单击 OK 按钮，在数据窗中显示变量 birthday1 的值，即每个同学生生日中的月份。见图 5-10。结果保存在数据文件是 data05-05 中。

## 习 题 5

1. 上网查询，为什么在 SPSS 的时间运算总是与 1582 年 10 月 14 日 24:00:00 (10 月 15 日 0:00:00) 有关，即日期型数据转换成以这个时间点为起始点的秒数？
2. 定义一个日期变量，输入你们班级同学的生日，计算他们此时的年龄。
3. 计算你们班同学中几个人的生日在 10 月？

# 第6章 构建表格

## 6.1 自定义表格

### 6.1.1 自定义表格的概念

一个好的表格能使统计资料系统化、层次化和条理化。在 SPSS 中的自制表格程序—Custom Tables 可以产生 1、2、3 维表格，各维度可用单个变量或变量组合来定义。在每一维中（行、列和层），可以叠放多重变量使之成为复合表，并为嵌套变量建立子表。

#### 1. 表格的组成

表格通常由行和列交叉组成。根据表格行、列中变量分层的多少，可将表格分成简单和复杂两种。最简单的表格形式可由单列、两行或单行、两列组成。表 6-1 是一张较复杂的表格，它由男、女对婚姻幸福程度感受的同结构的两个单表上下叠加而成。

表格通常包括以下几个部分。

(1) 表头。一般位于表格的上方，简明地描述表格所反映的中心内容。

(2) 行变量。定义表格的行，在表格左边由上到下排列的变量称为行变量。

(3) 列变量。定义表格的列，在表格上面横向排列的变量称为列变量。

(4) 单元格。由表格的行和列的交叉点形成。

(5) 表格的实体部分。由所有单元格组成，包括由表格总计、总和、平均数、百分比等基本信息。

(6) 角注位于右上角，脚注位于表格的下方。简要说明表格的组成、生成的日期、时间或其他需要特别的声明。

#### 2. 表格的结构类型

表 6-1 不同性别多子女人群对婚姻幸福程度的感受

				各组分类中子女的数量			
				0	1~2	3~4	>=5
				列%	列%	列%	列%
性别	男	婚姻幸福程度	很幸福	36.2	64.9	62.2	71.4
			中等幸福	32.9	33.0	35.2	26.2
			不太幸福	3.9	2.1	2.4	2.4
	女	婚姻幸福程度	很幸福	69.2	63.9	61.6	57.6
			中等幸福	28.6	33.2	35.5	23.3
			不太幸福	2.2	2.9	2.8	9.1

根据表格中变量的位置、作用可将表格分为以下几种类型：

- ① 简单表格，它是由一个变量和若干统计汇总指标组成，变量可以在行或在列上；
- ② 简单交叉表，即在行、列上均设置了一个变量，包括各变量的分类、统计汇总指标；
- ③ 堆栈式表格，即在行（或列）上有并列的两个以上变量；
- ④ 嵌套表格，即同行或同列上由不同层次上的变量组成的表格；
- ⑤ 分层表格，按某变量的分类分别形成表格，每层一个表。见图 6-1。

		Count
Gender	Male	1232
	Female	1600
	Less than 25	242
	25 to 34	627
	35 to 44	679
	45 to 54	481
	55 to 64	320
	65 or older	479

		Count
Age category	Less than 25	108
	Female	134
	25 to 34	276
	Male	351
	35 to 44	309
	Female	370
	45 to 54	221
	Male	260
	55 to 64	136
	Female	184
	65 or older	178
	Male	301
	Female	301

		Count
Gender Female		
Age category	Less than 25	134
	25 to 34	351
	35 to 44	370
Gender Male		
Age category	Less than 25	108
	25 to 34	276
	35 to 44	309
	45 to 54	221
	55 to 64	136
	65 or older	178

图 6-1 表格结构：堆栈、嵌套、分层表格列举

### 3. 变量的测度方法与 Custom Tables 能自动识别的测度标准

在 SPSS 中，测度方法共有三种，即 Ordinal 顺序变量、Nominal 名义变量和 Scale 尺度变量。前两者是分类变量，可以定义表格的行、列和层。默认的汇总统计指标是计数；尺度变量一般被汇总在分类变量的类别里，默认的汇总统计指标是算术平均数。在没有使用分类变量定义组别时，还可以用尺度变量本身来汇总尺度变量。它主要用于多重尺度变量的分层汇总。

### 4. Multiple Response Sets（多重应答集）

当对一个问题答案不止一个时，往往建立多个变量记录这些答案的数据。通过定义，多个变量构成一个多重应答集。Custom Tables 支持包含多重应答集表格，但 Custom Tables 不支持对多重应答集进行显著性检验。

## 6.1.2 自定义表格的操作

1. 按 Analyze→Tables→Custom Tables 顺序单击菜单项打开图 6-2 对话框。

在进入制表工作之前，应先定义好变量的各种属性，尤其是变量类型、变量标签、值标签、测定标准等。第一次启动如图 6-2 所示的 Custom Tables 对话框，单击 Define Variable Properties 按钮，在其后的对话窗口中将变量属性定义完全。

或者选择 Don't show this dialog again 和 OK 按钮，关闭这一对话框，进入 Table 主对话

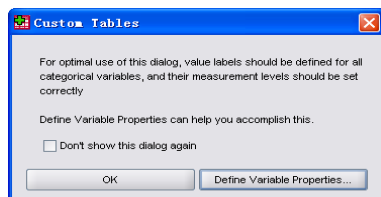


图 6-2 Custom Tables 预备对话框

框，见图 6-3。有四个选项卡，分别是 Table、Titles、Test Statistics 和 Options。

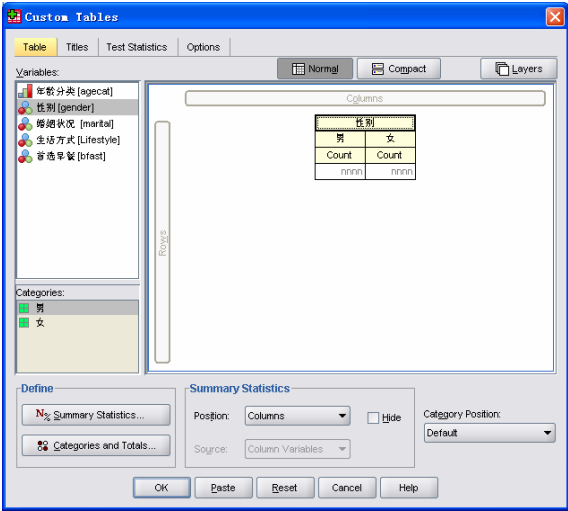


图 6-3 Custom Tables 的 Table 对话框

时，在探究窗口中显示默认的统计量名称 Count。

如果拖曳到行或列中的是尺度变量，只在探究窗口的行（或列）中显示变量名和默认的统计量名称 Mean。单击 OK 按钮，所建表格显示在输出窗口中。

3. 各种结构的表格

(1) 如要选择多个变量，并且将它们一起拖曳和置于探究窗口的行（或列）中，则所选择变量各分类在行（或列）中是并排的，生成堆栈表格。

(2) 如果将变量拖到已在窗口中的行变量左面（或右面），形成行嵌套，左面变量在外层。如果将变量拖至窗口中列变量的上方（或下方），形成列嵌套，上面的变量在外层。

(3) 表格的分层。当拖曳一个变量到探究窗口右上方 Layer 图标处，该变量被定义为层变量。无论测度方法是有序的还是标称的分类变量都可以作为层变量。层变量的每个分类的数据按探究窗口中设置的行、列变量构成一个表格，称为一层，每层结构相同。层变量数及其分类数决定表格数量，要慎重指定层变量。层变量之间是并列关系，层数等于各层变量分类数之和，例如 3 个 3 类分层变量共 9 层；层变量之间是嵌套关系，层数是各变量分类数之积，例如 3 个 3 类变量生成 27 层，即 27 个表格。

(4) 从探究窗口中选择一个变量，按 Delete 键或将其拖曳到窗口外即删除该变量。

2. 简单的制表操作过程

表格在 Table 选项卡中构建。

从变量列表中拖曳一个变量或多应答变量集到探究窗口行或列的区域，就可以建立一个最简单的表格。探究窗口显示将要建立的表格的预览。不显示实际数据值，而是显示所设计表格的轮廓。

如果拖曳到行或列中的是分类变量，变量表下方 Categories 栏和探究窗口中显示该变量的各分类值。如果该分类变量没有定义值标签，无论实际有几类 Categories 栏和探究窗口都只显示两个类别：Category 1 和 Category 2。与此同

## 6.2 汇总、统计指标与统计检验

### 6.2.1 统计指标与汇总项

在 Table 选项卡中，每定义一个行变量或列变量，就可以马上定义要在表格中出现的该变量的统计指标和汇总项。

#### 1. Defines 定义栏

在 Defines 栏中有两个选项，在探究窗口选择行、列上的尺度变量时，激活 Summary Statistics 选项；选择分类变量时，激活 Categories and Totals 选项。

#### (1) Summary Statistics 汇总统计指标

在探究窗口选择尺度变量，单击 Summary Statistics 按钮，进入定义这个变量的统计量的对话框，见图 6-4。在该窗口中，每选择一个统计量，送入 Display 表中，便可以在 Display 表中编辑这个统计量的标签 Label，从 Format 的下拉列表中选取统计指标的



图 6-4 综合统计量对话框

显示格式，输入小数位数 Decimal。单击右边的上下箭头改变当前统计指标的显示顺序。

单击 Apply to Selection 按钮，只对当前选择的变量计算指定的汇总统计指标。单击 Apply to All 按钮，对探究窗口所有同类型的变量都计算指定的汇总统计指标。

汇总统计指标的有效性取决于所选择变量的测定标准和变量的选择顺序。对嵌套表格来说，对嵌套在最内层的变量有效。分层表格是对每层中的指定变量有效。

① 适合于所有变量的汇总统计指标见表 6-2。

表6-2 适合所有变量的汇总指标

指 标	计算内容描述	默认标签	默认格式
COUNT*	各类中样品的数量。它对分类和多重应答变量是默认的	Count	Count
ROWPCT.COUNT	基于单元格计数的行百分比。在子表格内部计算	Row%	Percent
COLPCT.COUNT	基于单元格计数的列百分比。在子表格内部计算	Column%	Percent
TABLEPCT.COUNT	基于单元格计数的表格百分比	Table%	Percent
SUBTABLEPCT.COUNT	基于单元格计数的子表格百分比	Subtable%	Percent
LAYERPCT.COUNT	基于单元格计数的层百分比。未定义层时同表格百分比	Layer%	Percent
LAYERROWPCT.COUNT	基于单元格计数的行百分比。整行（即，子表）的百分比总和为100%	Layer Row%	Percent
LAYERCOLPCT.COUNT	基于单元格计数的列百分比。整列（即，子表）的百分比总和为100%	Layer Col%	Percent

（续表）

指 标	计算内容描述	默认标签	默认格式
ROWPCT.VALIDN	基于有效计数的行百分比	Row Valid N%	Percent
COLPCT.VALIDN	基于有效计数的列百分比	Col Valid N%	Percent
TABLEPCT.VALIDN	基于有效计数的表格百分比	Table Valid N%	Percent
SUBTABLEPCT.VALIDN	基于有效计数的子表格百分比	Subtable Valid N%	Percent
LAYERPCT.VALIDN	基于有效计数的层百分比	Layer Valid N%	Percent
LAYERROWPCT.VALIDN	基于有效计数的行百分比。整行的百分比和为100%	Layer Row Valid N%	Percent
LAYERCOLPCT.VALIDN	基于有效计数的列百分比。整列的百分比和为100%	Layer Col Valid N%	Percent
ROWPCT.TOTALN	基于总计数的行百分比，包括读者和系统缺失值	Row Total N%	Percent
COLPCT.TOTALN	基于总计数的列百分比，包括读者和系统缺失值	Column Total N%	Percent
TABLEPCT.TOTALN	基于总计数的表格百分比，包括读者和系统缺失值	Table Total N%	Percent
SUBTABLEPCT.TOTALN	基于总计数的子表格百分比，包括读者和系统缺失值	Subtable Total N%	Percent
LAYERPCT.TOTALN	基于总计数的层百分比，包括读者和系统缺失值	Layer Total N%	Percent
LAYERROWPCT.TOTALN	基于总计数的行百分比，包括读者和系统缺失值。整行的百分比和为100%	Layer Row Total N%	Percent
LAYERCOLPCT.TOTALN	基于总计数的列百分比，包括读者和系统缺失值。整列的百分比和为100%	Layer Col Total N%	Percent

\*在美国英语体系中这是默认的。后缀.COUNT 可以从计算基于单元格的百分比中省略。因而，ROWPCT等于ROWPCT.COUNT。

② 适合于分类变量的汇总统计指标

- Count，各单元格中样品的数量，或多重应答集中应答的数量。
- Unweighted Count，表格的每个单元格中样品的未加权的数量。
- Column percentages，列百分数。子表每列的百分数的和为 100%。仅当有分类行变量时，列百分数才是有效的。
- Row percentages，行百分数。子表的每行的百分数的和为 100%。仅当有分类列变量时，行百分数才是有效的。
- Layer Row and Layer Column percentages，每层中行百分比和列的百分比。嵌套表中，各子表的行或列百分数的总和为 100%。分层表格每层中所有嵌套子表的行或列的百分数总和为 100%。
- Layer percentages，每层内的百分数。作为简单百分数，当前可见层里单元格的百分数总和为 100%。如果没有层变量，则它等于同表格百分数。
- Table percentages，表中各单元格的百分数建立在整个表格基础上。所有单元格的百分数是建立在样品总数的基础上的，并且整个表格百分数的总和为 100%。
- Subtable percentages，子表中每个单元格的百分数建立在子表基础上。子表中，所有单元格的百分数建立在子表内相同样品总数的基础上，并且在子表内单元格的百分数

总和为 100%。在嵌套表中，在最里面嵌套水平之前的变量定义子表格。百分数受计算它们的基数（分母）的影响，并且选项的数量决定基数。基于样品、应答或计数，多重应答集可以有百分数。

由层变量定义的各层表格被当作独立的表格处理。在各层表格内，各层 Layer Row 的总和、各层 Layer Column 的总和以及每层 Table 百分数的总和都为 100%。

③ 适合尺度变量及分类自定义合计的主要统计指标见表 6-3。对分类变量还能求部分和及自定义求总和。表格默认包括总计或小计。

④ 适合多重应答集的统计指标见表 6-4。

表6-3 适合尺度变量、总计和小计的主要统计指标

指 标	描 述	默认标签	默认形式
MAXIMUM	最大值	Maximum	General
MEAN	算术平均数。默认尺度变量	Mean	General
MEDIAN	中位数	Median	General
MINIMUM	最小值	Minimum	General
MISSING	缺失值合计（用户和系统的缺失值）	Missing	General
MODE	众数。如果有结（众数相同），则显示最小的值	Mode	General
PTILE	百分位数。取0至100间的一个数值作为需要的参数。PTILE在SPSS Tables中是用同样的PTILE来计算的。注意，在SPSS Tables中默认的百分位数的方法是HPTILE	Percentile ####.##	General
RANGE	两极差	Range	General
SEMEAN	标准误	Std Error of Mean	General
STDDEV	标准差	Std Deviation	General
SUM	数值总和	Sum	General
TOTALN	非缺失值、用户缺失值和系统缺失值的合计。由CATEGORIES子命令中隐含的有效值不计数	Total N	Count
VALIDN	非缺失值合计	Valid N	Count
VARIANCE	方差	Variance	General
ROWPCT.SUM	基于总和的行百分比	Row Sum%	Percent
COLPCT.SUM	基于总和的列百分比	Column Sum%	Percent
TABLEPCT.SUM	基于总和的表格百分比	Table Sum%	Percent
SUBTABLEPCT.SUM	基于总和的子表格百分比	Subtable Sum%	Percent
LAYERPCT.SUM	基于总和的层百分比。	Layer Sum%	Percent
LAYERROWPCT.SUM	基于总和的行百分比。整行的百分比总和为100%	Layer Row Sum%	Percent
LAYERCOLPCT.SUM	基于总和的列百分比。整列的百分比总和为100%	Layer Column Sum%	Percent

表6-4 适合多重应答集的主要统计指标

指 标	计算内容描述	默认标签	默认形式
RESPONSES	回答合计	Responses	Count
ROWPCT. RESPONSES	行百分比，回答合计是分子。回答的总计是分母	Row Responses%	Percent
COLPCT. RESPONSES	列百分比，回答合计是分子。回答的总计是分母	Column Responses%	Percent
TABLEPCT. RESPONSES	表格百分比，回答合计是分子。回答的总计是分母	Table Responses%	Percent
SUBTABLEPCT. RESPONSES	子表格百分比，回答合计是分子。回答的总计是分母	Subtable Responses%	Percent
LAYERPCT. RESPONSES	层百分比，回答合计是分子。回答的总计是分母	Layer Responses%	Percent



(续表)

指 标	计算内容描述	默认标签	默认形式
LAYERROWPCT.RESPONSES	层中行百分比，回答合计是分子。回答的总计是分母，子表格中整行的百分比和为100%	Layer Row Responses %	Percent
LAYERCOLPCT.RESPONSES	层中列百分比，回答合计是分子。回答的总计是分母，子表格中整列的百分比和为100%	Layer Column Responses%	Percent
ROWPCT.RESPONSES.COUNT	行百分比：行的回答合计是分子，总计为分母	Row Responses% (Base: Count)	Percent
COLPCT.RESPONSES.COUNT	列百分比：列的回答合计是分子，总计为分母	Column Responses% (Base: Count)	Percent
TABLEPCT.RESPONSES.COUNT	表格百分比：表格的回答合计是分子，总计为分母	Table Responses% (Base: Count)	Percent
RESPONSES	分母	(Base: Responses)	
SUBTABLEPCT.COUNT.RESPONSES	子表百分比：子表中的合计是分子，回答的总计是分母	Subtable Count% (Base: Responses)	Percent
LAYERPCT.COUNT.RESPONSES	层百分比：层中的合计是分子，回答的总计是分母	Layer Count% (Base: Responses)	Percent
LAYERROWPCT.COUNT.RESPONSES	行百分比：行中的合计是分子，回答的总计是分母。整行（即子表格）的百分比和为100%	Layer Row Count% (Base: Responses)	Percent
LAYERCOLPCT.COUNT.RESPONSES	列百分比：列中的合计是分子，回答的总计是分母。整列（即子表格）的百分比和为100%	Layer Column Count% (Base: Responses)	Percent
SUBTABLEPCT.RESPONSES.COUNT	子表格百分比：子表的回答合计是分子，总计为分母	Subtable Responses% (Base: Count)	Percent
LAYERPCT.RESPONSES.COUNT	计算层百分比：层的回答合计是分子，总计为分母	Layer Responses% (Base: Count)	Percent
LAYERROWPCT.RESPONSES.COUNT	计算行百分比：行的回答合计是分子，总计为分母。整行（即子表格）的百分比和为100%	Layer Row Responses% (Base: Count)	Percent
LAYERCOLPCT.RESPONSES.COUNT	计算列百分比：列的回答合计是分子，总计为分母。整行（即子表格）的百分比和为100%	Layer Column Responses% (Base: Count)	Percent
ROWPCT.COUNT.RESPONSES	行百分比：行的回答合计是分子，回答的总计是分母	Row Count% (Base: Responses)	Percent
COLPCT.COUNT.RESPONSES	列百分比：列中的回答合计是分子，回答的总计是分母	Column Count% (Base: Responses)	Percent

⑤ Custom Total and Subtotal Summary Statistics。使用 Summary Statistics 对话框，在表格的 Totals and Subtotals 区域，可以显示除 Totals 之外的其他统计指标。

(2) Categories and Totals 分类与总计

当拖入一个分类变量或多重应答集到探究窗口后，单击 Define 栏中的 Categories and Totals，进入相应的对话框，见图 6-5。

① Display 和 Exclude。在 Display 中显示生成的表格中分类变量的所有值和值标签。单击上、下箭头，可移动分类值在表中的位置。若要生成表格不包括某个分类值，将其移入 Exclude 栏中。

② Sort Categories。在该栏中, By 的下拉列表中选择排序依据, 可以选择按值 Value、值标签 Label 或单元格频数 Cell Count 排序; 在 Order 的下拉列表中选择 Ascending 或 Descending 决定排序是升序还是降序。

③ Subtotal 小计。如果需要对部分分类值计算小计, 在 Display 中选择一个分类值, 单击 Insert 按钮, Subtotal 项插入到 Display 栏中, 并在 Value 列中显示小计的范围。

④ 选择 Show 栏中的 Total 总计。可以继续选择特殊分类总计的内容: Missing Value 缺失值、Empty categories 频数为 0 的分类、Other values found when data are scanned 对数据扫描时发现的其他值。

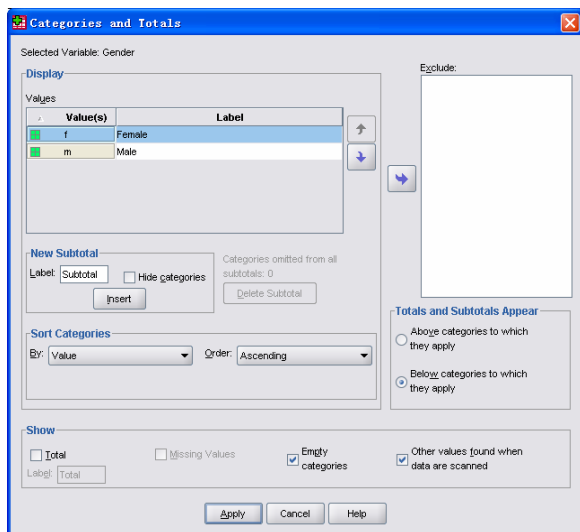


图 6-5 Categories and Totals 对话框

⑤ Totals and Subtotals Appear 总计和小计显示位置, 相对于被小计 (或被总计) 的所有分类值的位置: Above 上方, 或 Below 下方。

设置完成单击 Apply 按钮, 确定选择, 返回 Table 选项卡对话框。

⑥ 在 Table 对话框中 Summary Statistics 栏选择统计指标出现的位置和维度。

- Position 下拉列表中, 有两个选项: 选择 Columns, 统计指标出现在列中, 选择 Rows, 统计指标出现在行中。如果在表中不想出现统计指标, 选择 Hide。

- Source 下拉列表中, 可通过选择 Column Variables、Row Variables、Layer Variables 来改变汇总统计的维度。

- 在 Category Position 的下拉列表中, 选择 Default 按系统默认格式显示; Row Labels in Columns 或 Column Labels in Rows 分别为行分类标签显示在列中或列分类值标签显示在行中。

⑦ 界面方式的切换按钮在探究窗口上方, Normal 界面显示包括在表格中的所有行、列及分类变量的类别和汇总统计等内容。Compact 界面只显示表格中的变量名及其位置。

## 2. Layers (层)

单击探究窗口右上角的 Layers, 显示层变量列表。多重应答集被当作分类变量列出。如果有两个以上的层变量, 可以在下面两个选项中, 确定层的构建方式:

- Show each category as a layer 每个层变量的每个分类为一层。层数为分类数总和。
- Show each combination of the categories as a layer 每个层变量各分类组合为一层,

层数为各层变量分类数的乘积。

## 6.2.2 表格中的统计检验

在主对话框中单击 **Test Statistics** 选项卡，进入如图 6-6 的对话框。在其中，可以对自定义表格中的变量作不同的显著性统计检验。选项包括：

(1) **Tests of independence(Chi-square)**，独立性卡方检验。行和列中最少有一个分类变量的表格可以选择此项；还可以指定检验的  $\alpha$  水平。

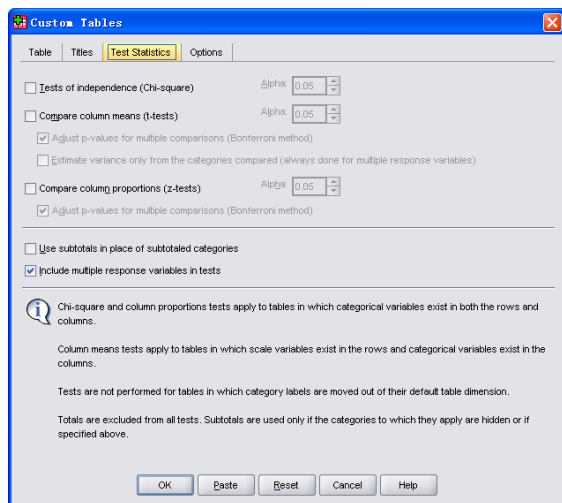


图 6-6 Test Statistics 选项卡

(2) **Compare column means (t-tests)**，列的均数相等检验。在列中至少有一个分类变量和在行中至少有一个尺度变量的表格可以选择此项。可以选择是否用邦弗伦尼（Bonferroni）方法校正检验的  $P$  值。还可指定检验的  $\alpha$  水平，该值应大于 0 和小于 1。

(3) **Compare column proportions (z-tests)**，列比率相等检验。行和列中至少存在一个分类变量的表格可以选择此项。可以选择是否用邦弗伦尼方法校正检验的  $P$  值。还可以指定检验的  $\alpha$  水平，该值应该大于 0 和小于 1。

## 6.3 标题与其他选项

### 6.3.1 定义表格标题

在表格自定义对话框中，单击 **Titles** 选项卡，如图 6-7。

(1) 在 **Title** 框中输入表格的标题。

(2) 在 **Caption** 框中输入脚注。

(3) 在 **Corner** 框中输入显示在表格左上角的说明文字。只有当定义行变量且当基准行维度标签已设置成 **Nested**（嵌套）时才显示。这不是默认表格外型的设置。

(4) 插入当前日期、时间的方法是将插入点光标移至 **Title**、**Caption** 栏中，单击选项卡上方的日期或时间图标。

(5) **Table Expression** 在 **Title** 对话框中插入此项标识，产生的表格相应位置显示各变量在表格中的作用： $+$  表示分层变量； $>$  表示嵌套变量；**BY** 表示交互变量或层变量。

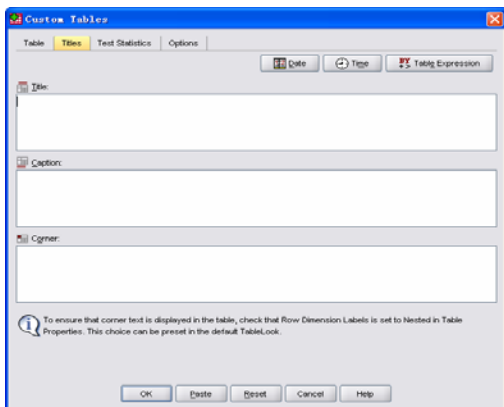


图 6-7 Titles 选项卡

### 6.3.2 定义表格选项

在自定义表格主对话框中单击 Options，可进入图 6-8 所示的选项卡。

(1) Data Cell Appearance 栏，定义空单元及无法进行统计计算的单元里显示什么。

① Empty Cells，对计数为 0 的单元能选择显示: Zero(0)、Blank（空格）或说明文字 Text。文本最大长度为 255 个字符。

② Statistics that cannot be computed 对指定的统计量不能计算时，例如分类里没有样品的均数，相应位置要显示的文字，字符数小于等于 255。默认值用圆点表示。

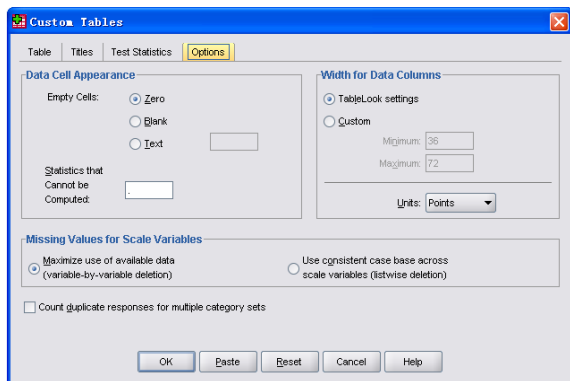


图 6-8 Options 选项卡

(2) Width for Data Columns 栏，定义数据列最小和（或）最大宽度。

① TableLook settings，使用默认的表格外观参数中的列宽度。

② Custom，读者指定最小和最大的列宽度和单位 Unit: Points（点）、Inches（英寸）或 Centimeters（厘米）。

(3) Missing Values for Scale Variables 栏，尺度变量的缺失值。对有两个或更多的尺度变量的表格定义计算尺度变量的统计指标时有关缺失数据的处理方法。

① Maximize use of available data (variable-by-variable deletion)，在默认表格中，主要统计指标包含每个尺度变量所有具有有效值的样品。

② Use consistent case base across scale variables(listwise deletion)，对表格里任何尺度变量而言，对其主要统计指标的计算时剔除所有带有缺失值的样品。

(4) Count duplicate responses for multiple category sets，对多重应答集中的变量，在默认情况下，不计重复应答的数量；选择此项，合计重复应答的数量。

6.4 自定义表格实例

【例】对不同性别、年龄、婚姻状况的生活方式和首选早餐的调查研究，数据编号为data06-01，操作方法如下。

- (1) 打开data06-01，按Analyze→Tables→Custom Tables顺序单击菜单项，进入Table主对话框。
- (2) 将变量gender拖曳到Rows中，将年龄段变量agecat拖曳并嵌套在gender变量下，将婚姻状况marita拖曳并嵌套在agecat变量下，将变量active（Lifestyle）、bfast首选早餐移到Columns中。
- (3) 单击Title按钮，进入Titles选项卡。在Title下面的编辑框中输入表头“不同性别、年龄、婚姻状况的生活方式和首选早餐的统计表”；将插入键移到Caption下面的编辑框中，单击Date和Time按钮。设置脚注为制表时的计算机系统的当前日期和时间。
- (4) 单击 Test Statistics 按钮，进入相应的对话框。选择 Tests of independence (Chi-square)项。对表格的列变量的分类间作独立性的卡方检验。
- (5) 单击OK按钮，运行程序，在输出窗口中得到如表6-5和表6-6所示的输出结果。

表6-5 频数分布表

不同性别、年龄、婚姻状况的生活方式和首选早餐的统计表										
性别 男 年龄 分类			婚姻状况			生活方式		首选早餐		
						不活动	活动	早餐吧	麦片	谷类
						Count	Count	Count	Count	Count
男	< 31	婚姻状况	未婚	18	26	25	0	19		
			已婚	14	27	15	0	26		
		31-45	未婚	10	14	13	2	9		
			已婚	33	40	26	12	35		
		46-60	未婚	5	13	5	7	6		
			已婚	60	35	16	37	42		
	> 60	婚姻状况	未婚	31	10	4	27	10		
			已婚	65	23	0	70	18		
		女	< 31	婚姻状况	未婚	15	33	27	1	20
					已婚	23	25	17	3	28
			31-45	婚姻状况	未婚	12	16	17	1	10
					已婚	34	47	34	9	38
	46-60	婚姻状况	未婚	19	5	7	10	7		
			已婚	55	39	11	43	40		
	> 60	婚姻状况	未婚	49	27	10	47	19		
			已婚	31	26	4	41	12		

2005-8-222:22:21

表6-6 皮尔逊卡方检验

Pearson Chi-Square Tests										生活方式	首选早餐
性别	男	< 31	婚姻状况	Chi-square	.414						3.487
				df	1						1
		31-45	婚姻状况	Sig.	.520						.062 <sup>a</sup>
				Chi-square	.092						2.802
		46-60	婚姻状况	df	1						2
				Sig.	.762						.246
	女	> 60	婚姻状况	Chi-square	7.752						1.395
				df	1						2
		31-45	婚姻状况	Sig.	.005 <sup>*</sup>						.498
				Chi-square	.045						9.482
		< 31	婚姻状况	df	1						2
				Sig.	.832						.009 <sup>a</sup>
	年龄分类	> 60	婚姻状况	Chi-square	2.788						4.606
				df	1						2
		31-45	婚姻状况	Sig.	.095						.100 <sup>a</sup>
				Chi-square	.007						3.443
		46-60	婚姻状况	df	1						2
				Sig.	.935						.179
	婚姻状况	> 60	婚姻状况	Chi-square	3.488						4.754
				df	1						2
		31-45	婚姻状况	Sig.	.062						.093
				Chi-square	1.383						1.885
		< 31	婚姻状况	df	1						2
				Sig.	.240						.390

Results are based on nonempty rows and columns in each innermost subtable.

<sup>a</sup>. The Chi-square statistic is significant at the 0.05 level.

<sup>a</sup>. More than 20% of cells in this subtable have expected cell counts less than 5. Chi-square results may be invalid.

表6-5反映不同性别、年龄、婚姻状况的生活方式和首选早餐的人数统计。

表6-6是皮尔逊卡方检验表。表中列出了不同性别、年龄、婚姻状况下的生活方式和首选早餐两个变量的各项间的皮尔逊卡方独立性检验，两个数据列中第一个数据是卡方值，第二个数据是自由度，第三个数据是原假设成立的显著性概率值，该值小于0.05说明在该性别、年龄、婚姻状况下的生活方式或首选早餐的各项间有差异。其余类推。

## 6.5 自定义表格的过程语句

### 1. 过程语句清单

#### CTABLES

```

/FORMAT MINCOLWIDTH={DEFAULT} {value } MAXCOLWIDTH={DEFAULT} {value }
      UNITS={POINTS} {INCHES} {CM } EMPTY= {ZERO} {BLANK } {'chars'} MISSING=
      {'.' } {'chars'}

/VLABELS VARIABLES= varlist
/DISPLAY= {DEFAULT} {NAME } {LABEL } {BOTH } {NONE }
/MRSETS COUNTDUPLICATES= {NO} {YES}
/SMISSING {VARIABLE} {LISTWISE}
/TABLE rows BY columns BY layers
/SLABELS POSITION= {COLUMN} {ROW } {LAYER } VISIBLE= {YES} {NO }
/CLABELS {AUTO} {ROWLABELS= {OPPOSITE} {LAYER} } {COLLABELS= {OPPOSITE}
      {LAYER} }

/CATEGORIES VARIABLES= varlist {[value, value, value...]}
      {ORDER= {A} {D} KEY= {VALUE} {LABEL} MISSING= {EXCLUDE} {INCLUDE} }
      {summary(varname)} TOTAL= {NO} {YES} LABEL= "label"
      POSITION= {AFTER} {BEFORE} EMPTY= {INCLUDE} {EXCLUDE}
      Explicit value lists can include SUBTOTAL= 'label', MISSING, and OTHERNM.
/TITLES CAPTION= ['text' 'text'...] CORNER= ['text' 'text'...] TITLE= ['text' 'text'...]
      Text can contain the symbols) DATE) TIME) TABLE
/SIGTEST TYPE= CHISQUARE ALPHA= {0.05} {significance level}
/COMPARETEST TYPE= {PROP} {MEAN} ALPHA= {0.05} {significance level}
      ADJUST= {BONFERRONI} {NONE} ORIGIN=COLUMN

```

各子命令中的加粗字符、数字是默认参数。

### 2. CTABLES语句

该语句调用自定义表格过程，至少包括一个TABLE子命令定义一个表格。全局的子命令FORMAT、VLABELS、MRSETS和SMISSING必须出现在第一个TABLE子命令前，

顺序任意。局部的子命令SLABELS、CLABELS、ATEGORIES、TITLES、SIGTEST和COMPARETEST跟在TABLE子命令之后，顺序任意，并立即对它之前的表格表达式起作用。一般地，如果子命令重复出现，最后一个有效。

除汇总指标名必须完整拼写外，所有关键字、属性值和外在的分类列表关键字可以用前三个字符的缩写形式。

3. TABLE子命令

TABLE子命令是必须使用的子命令。它指定表格结构，按行、列层顺序指定表格变量。变量间用BY连接，每个变量后面用中括号包含对该变量使用的格式、标签、汇总指标等。各维变量名及其后面中括号中的内容构成表格表达式。表格表达式一般规则如下：

(1) 同一维度变量之间使用连接符号表达结构，嵌套(>)、连接(+)。当连接符号连用时，用括号改变优先级。如 $a + b > c$ 表示先嵌套再连接； $(a + b) > c$ 表示先连接再嵌套。

(2) 在表格表达式中可以使用分类变量、尺度变量或多重应答集。多重应答集由命令MRSETS定义，以\$开头。尺度变量只能在一个维度中使用，不能嵌套。

(3) 使用尺度变量的维度作为统计维度，可以指定尺度变量的所有统计指标。只能为最低嵌套水平的分类变量指定汇总指标。如果样品加权是有效的，为区分加权指标与未加权指标，可以在指标名前加“U”表明未加权指标，如UCOUNT是未加权计数。

(4) 汇总格式很多，控制输出或打印形式有COMMAw.d、DOLLARw.d、Fw.d、NEGPARENw.d、NEQUALw.d、PARENw.d、PCTw.d、PCTPARENw.d、DOTw.d、CCA...CCEw.d、Nw.d、Ew.d和所有DATE的格式，对应所有变量的格式。其中w为宽度，d为小数位数。关键字与所有变量格式对应。在此不再赘述。

系统默认的打印格式，对计数count使用整数。百分比保留一位小数，加“%”符号。另外的附加格式见表6-7。

(5) 各类型变量的汇总指标VALIDN、COUNT和TOTALN对缺失值的处理见表6-8。

4. 其他子命令

(1) SLABELS子命令用“POSITION=”连接关键字COLUMN、ROW、LAYER、VISIBLE= YES或NO，定义汇总统计指标显示在列、行、层和是否显示汇总标签。默认汇总出现在列中并可见到标签。

表6-7 汇总的附加格式

格 式	描 述	例 子
NEGPARENw.d	圆括弧出现在负数的两边	-1234.567用格式NEGPAREN9.2产生(-1234.57)
NEQUALw.d	在数字前加“N=”	1234.567用格式NEQUAL9.2产生N=1234.57
PARENw.d	将数字加上括弧	1234.567用格式PAREN8.2产生(1234.57)
PCTPARENw.d	数值后接%符号，再用圆括弧将它们括起来	1234.567用格式PCTPAREN10.2产生(1234.57%)

表6-8 包含或不包含在汇总里的数值

变量和值类型		VALIDN	COUNT	TOTALN
分类变量: 显示有效值 多重分类集: 至少显示一个有效值	多重二分集: 至少有一个“true”值 尺度变量: 有效值	包含	包含	包含
分类变量: 包含用户缺失值 尺度变量: 用户缺失值或系统缺失值	多重分类集: 所有值包含用户缺失值	不包含	包含	包含
分类变量: 不包含用户缺失值或系统缺失值 多重二分集: 所有值是false 多重分类集: 所有值是不包含用户缺失值、系统缺失值或不包含无效值, 但至少一个值包含有效值		不包含	不包含	包含
分类变量: 不包含有效值 多重二分集: 所有值不包含有效值		不包含	不包含	不包含

(2) CLABELS子命令用“ROWLABELS=”和“COLLABELS=”定义行、列分类标签的位置。默认分类标签出现在它们所属的变量下面。用关键字OPPOSITE和LAYER表示标签显示在另一个维度中或在层上。如果两个维度都要显示标签，只有一个维度的行（或列）标签可以移动。

(3) CATEGORIES 子命令指定分类变量在表格中出现哪些值，以及在行和列中的次序、缺失值的显示和隐藏，以及总计和小计的计算。在该子命令中，最少指定一个变量并在后面的括号中用逗号分开列出要出现在生成表格中的该变量的值，可以使用关键词ALL、LO、HI和  $n_1$  THRU  $n_2$  的关键词和数字表示所有值、最低值、最高值、从 $n_1$ 到 $n_2$ 值。

① ORDER= 后面指定这些值出现的顺序，A表示升序，D表示降序；

② KEY=后面用VALUE要求按变量值排序，LABEL表明按值标签排序；

③ MISSING=后面用EXCLUDE表示不出现缺失值，INCLUDE表示可出现缺失值；

④ TOTAL=后面用YES表示表格中要有该变量的总计，NO说明不要总计。LABEL 后面用引号中的字符串给出总计标签；POSITION=后面用AFTER或BEFORE要求总计出现在被总计变量的分类之后，或是在它之前；可以在多个维度中要求总计。

⑤ EMPTY= 后面用EXCLUDE要求表中不出现空单元，INCLUDE表示包括空单元。

⑥ SUBTOTAL 要求对变量值做小计。关键字SUBTOTAL之前列出要小计的分类，等号后面用引号中的字符串给出小计标签，POSITION=BEFORE指定对后面的分类值进行小计。

⑦ OTHERNM 关键字是指在列表中没有明确给出的全部非缺失值。它可以放在列表中的任何地方，提交的值按升序出现。

一个多重二分变量集中，变量的顺序按集中的变量名排列。变量名不用装在引号里。

(4) TITLES子命令用TITLES=、CAPTION=、CORNER=各自后面引号中的字符串给出要在表格中出现的标题、脚注和角文本。角文本出现在表格的行标题上方和列标题的旁边。标题和脚注都可以指定多个引号括起的字符串，每个占一行。脚注、角文本或标题里还可以用“DATE”、“TIME”、“TABLE”要求显示当前日期、时间和表格描述。



这三个关键字必须大写。表格描述自动产生,由表格表达式、统计说明和TABLE子命令组成对表格的描述。如果定义了变量标签,则用变量标签代替表达式里的变量名。

(5) SIGTEST子命令在行和列中都最少包括一个分类变量时,用TYPE= CHISQUARE指定进行独立性卡方检验。ALPHA=指定检验的显著性水平,默认值为0.05。

(6) COMPARETEST子命令指定率以及均数的成对比较

TYPE=后面指定成对比较类型,关键字PROP、MEAN分别指定比较分类变量的率间差异、尺度变量均数间的差异;ALPHA=指定检验的显著性水平。默认值为0.05,指定值必须大于0小于1。ADJUST=指定进行多重比较的校正P值的方法NONE或BONFERRONI。默认使用BONFERRONI修正。ORIGIN=COLUMN指定对列均数进行趋势比较。

(7) FORMAT子命令控制表格的外观

① MINCOLWIDTH=指定列的最小宽度。其值必须小于等于设置的最大列宽。

② MAXCOLWIDTH=指定表格列的最大宽度。其值须大于等于设置的最小列宽值。

③ UNITS=指定列宽度值的单位,默认值为POINTS,也能指定INCHES或CM。

④ EMPTY=指定计数或百分比为0时的填充字符。ZERO、BLANK指定单元格中显示0或空白。也可在引号中指定一个字符串。当字符串长度超过单元格的宽度时,多余部分被删除。如果FORMAT EMPTY = BLANK,计数为0的和未定义统计的单元格显示相同。

⑤ MISSING=指定当统计指标不能计算时,填充单元格的字符。默认显示圆点。也可以在引号中指定一个字符串。若字符串的长度超过单元格的宽度,多余部分将被删除。

(8) VLABELS 子命令指定表格中的变量,与 DISPLAY 子命令配合说明在表格中对指定变量显示变量名还是标签或两者都显示或都不显示。用ALL或Varname TO Varname指定表格中的全部变量或某些变量,后者与数据文件里变量的顺序有关;需慎重使用。

DISPLAY子命令对VARIABLES子命令中指定的变量,用等号后面的关键字DEFAULT、NAME、LABEL和BOTH表示在表格中使用SET TVARS设置、只显示变量名、只显示变量标签或变量名和标签都显示。选择NONE隐藏变量名和标签。

(9) SMISSING子命令

处理尺度变量的缺失值。如果在表格里指定的尺度变量超过一个,在SMISSING子命令后面,使用VARIABLE指定要求仅剔除有缺失值的变量;用LISTWISE指定只要有一个变量有缺失值,变量汇总统计中剔除有缺失值的整个观测量,以保证所有尺度变量的汇总统计建立在相同的样品集基础上。对每一张基础表都删除带有缺失值的观测量。

(10) MRSETS子命令

多重应答集中,有的问题允许有相同的答案,例如问被访者拥有的汽车牌子,可能一个人有两辆相同牌子的汽车。有的多项选择题不允许多个相同的答案,例如要求回答阅读杂志的名称。MRSETS子命令允许指定重复的是否要计数。系统默认NO,即重复的不计数。选择YES,则重复的计数。MRSETS指令只适用于RESPONSES和基于RESPONSES的百分比。它不影响合计,合计中不包含重复出现的应答。

## 习 题 6

1. 用数据文件 data06-02 制表，表明不同性别、不同民族的平均工资，以职务等级作为层变量，自己定义表格标题。

2. 用数据文件 data06-02，将受教育程度重新分段编码：≤8 年的编码为 1；9 年～12 年编码为 2；13～16 年的编码为 3，17 年以上的编码为 4。制表，表明不同受教育年限的各种职务的人数；不同受教育年限的各种职务的平均初始工资。性别做层变量。

## 第 7 章 基本统计分析

在 SPSS 的 Analyze 菜单下的 Descriptive Statistics 中包括了一系列基本统计分析过程。常用的是频数分布表、描述统计量、交叉列联表及其独立性检验以及探索分析等。

当我们得到了审核无误的原始数据后，需要认识数据、了解数据特征及检查数据分布，以便对数据进行进一步处理和确定分析方法。

针对离散变量（也称分类变量），主要使用频数分布分析：Frequencies 过程可以通过一维的简单频数分布分析认识单一变量的分布；通过 Crosstabs 过程可以完成二维以上的交叉表、分层交叉表的分析，从而认识变量间的关系并进行变量间的独立性检验。

对于连续变量（即尺度测度的变量），可以通过 Descriptives 过程计算描述统计量认识变量值的集中性特征和波动性特征。反映数据集中性特征的统计指标称为集中趋势指标，如均值、中位数和众数等；反映数据波动性特征的统计指标称为离中趋势指标，如全距（也称为极差）、平均差、方差和标准差等。另外还可以计算偏度、峰度指标对变量的分布进行描述。

正态分布是连续型变量的概率分布之一，在统计学中非常重要。它是中间分布的频数多，两边分布的频数少，以均数为中心，左右对称的频数分布。变量是否服从正态分布是选择统计方法的前提条件。例如正态或近似正态分布的变量可以使用均值、标准差简要描述，而对非正态分布变量就要用最大值、最小值、极差、中位数、众数等来描述。又比如两组均值比较的 T 检验其前提条件是变量服从正态分布，否则应该使用非参数检验；在 SPSS 中可通过 Explore 过程或 P-P 图和 Q-Q 图对变量是否服从正态分布进行验证。

SPSS 的描述统计过程中还包括比率分析 Ratio 过程，它是用于对两个变量值比率变化的描述分析，适用于尺度变量。

### 7.1 频数分布分析

利用频数分布表可以对数据按组进行归类整理，形成各变量不同水平的频数分布表和频数分布图，从而对各变量的数据分布特征有一个基本认识。也可以通过该过程完成对数据的检查。

#### 7.1.1 频数分布分析过程

1. 建立或打开数据文件后，按 Analyze→Descriptive Statistics→Frequencies 顺序单击

菜单项，打开如图 7-1 所示的对话框。

2. 在源变量框中选择一个或多个变量，将其送入 Variable(s)框中。
3. 选中 Display frequency tables，要求输出频数分布表。
4. 单击 Statistics 按钮，打开如图 7-2 所示对话框，选择要求输出的统计量。

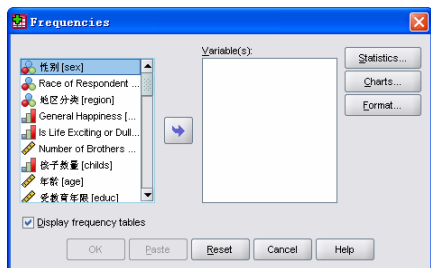


图 7-1 频数分布主对话框

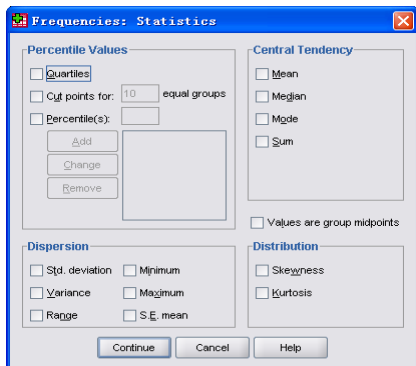


图 7-2 选择输出统计量对话框

(1) Percentile Values 栏，指定百分位数。

① Quartiles，输出四分位数，即第 25、50、75 百分位数。

② Cut points for \_\_\_equal groups，输出等分点的百分位数。在参数框中输入为 2~100 间的整数。例如，如果输入了 5，输出第 20、40、60、80 百分位数。

③ Percentile(s)，自定义百分位数。在参数框中输入 0~100 之间的数值，单击 Add 按钮。可多次重复此操作，可指定输出多个百分位数。要剔除已定义的百分位数，只需选择它，单击 Remove 按钮。

(2) Dispersion 栏，离散趋势统计量栏。

① Std. deviation，输出标准差。

② Variance，输出方差。

③ Minimum、Maximum 和 Range，输出的分别是最小值、最大值和全距。

④ S.E. mean，输出均数的标准误。

(3) Central Tendency 栏，指定集中趋势统计量。有以下复选项：Mean（均值）、Median（中位数）、Mode（众数）和 Sum（算术和）。

(4) Distribution 栏，指定描述数据分布的指标。

① Skewness，输出偏度值，同时显示偏度的标准误。偏度值位 0 说明变量是对称的。

② Kurtosis，输出峰度值及其标准误。峰度值为 0 说明变量是正态的。

(5) 选中 Values are group midpoints，在计算百分位数值和中位数时，假设数据已经分组，用各组的组中值代表各组数据。

5. 单击 Charts 按钮，展开如图 7-3 所示的 Frequencies: Charts 对话框。在其中设置

统计图的类型及坐标轴等。

(1) Chart Type 栏，选择统计图类型。

① None，不输出统计图，系统默认状态。

② Bar charts，输出条形图，各条高度代表变量各分类的频数或百分比。不显示频数为 0 的分类。条形图适用于分类变量。

③ Pie charts，输出饼图（或称圆图）不显示频数为 0 的类。饼图适用于分类变量。

④ Histograms，直方图，此图仅适用于连续型变量。选择此项后，选 With normal curve，显示的直方图中带有正态曲线。

(2) Chart Values 栏在选择了条形图和饼图栏后生效。

① Frequencies，直方图纵轴表示频数，饼图中的每个扇形表示该部分观测值的频数。

② Percentages，直方图纵轴表示百分比，饼图的每个扇形为观测值频数占总数的百分比。

6. 在主对话框中单击 Format 按钮，打开如图 7-4 所示的格式对话框。在其中设置频数分布表输出格式。

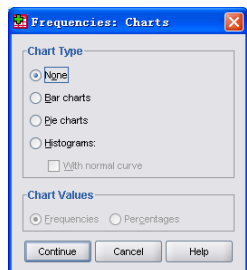


图 7-3 统计图类型选择对话框

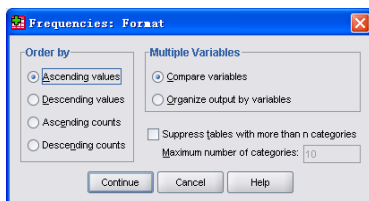


图 7-4 频数分析：格式对话框

(1) Order by 栏，设置频数分布表顺序，在选择了 Display frequency tables 后生效。

① Ascending values，按变量值的升序排列，这是默认的排列方式。

② Descending values，按变量值的降序排列。

③ Ascending counts，按变量值频数的升序排列。

④ Descending counts，按变量值频数的降序排列。

如果设置了直方图或百分位数，频数分布表按变量值升序排列，而忽略用户的设置。

(2) Multiple Variables 栏，选择多变量输出表格。

① Compare variables，所有变量的频数表集中输出。

② Organize output by variables，每一个变量单独输出一个频数表。

(3) Suppress tables with more than n categories，控制频数表输出分类数。如果变量值的个数太多，占用空间，此时可以压缩它。默认值为 10，即如果变量值的个数大于 10，则不输出相应的频数分布表。

### 7.1.2 频数分布分析实例

【例1】数据 data07-01 为 1991 年美国社会调查数据。变量: race (种族)、happy (幸福感)。要求编制 race、happy 变量的频数分布表。

操作步骤:

打开数据文件 data07-01, 按 Analyze→Descriptive Statistics→Frequencies 顺序打开 Frequencies 主对话框, 在左侧框中分别选中 race、happy 变量, 单击中间的箭头按钮, 将其送入 Variable(s)框中, 确认选中 Display frequency tables, 要求输出频数分布表, 单击 OK 按钮, 提交运行。

输出结果如见表 7-1 和表 7-2。

表 7-1 种族变量的频数分布表

种族				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 白人	1264	83.3	83.3	83.3
黑人	204	13.4	13.4	96.8
其他	49	3.2	3.2	100.0
Total	1517	100.0	100.0	

表 7-2 幸福感变量的频数分布表

幸福感				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 非常幸福	467	30.8	31.1	31.1
比较幸福	872	57.5	58.0	89.0
不太幸福	165	10.9	11.0	100.0
Total	1504	99.1	100.0	
Missing NA	13	9		
Total	1517	100.0		

结果解释:

从表 7-1 中可看到, 白种人 1264 人, 占 83.3%; 黑种人 204 人, 占 13.4%; 其他是 49 人, 占 3.2%。没有缺失值。从结果看, 本次调查中大部分是白种人。

从表 7-2 中可看到, 有 13 个缺失值。被调查者中有 467 人感到非常幸福, 有效百分比是 31.1%; 872 人感到比较幸福, 有效百分比是 58.0%; 165 人感到不太幸福, 有效百分比是 11.0%。从结果看, 有一半的被调查者感到比较幸福, 有 89% 的被调查者是幸福的 (包括比较幸福和非常幸福)。

说明: 本例中计算的两个变量中, race 变量的测度类型是 Nominal, happy 变量的测度类型是 Ordinal, 这两种类型的数据适合使用 Frequencies 过程进行频数分布分析。

【例2】我们仍然使用数据文件 data07-01。变量: age 年龄、educ 最高受教育年限。要求分析年龄和最高受教育年限的分布特征和描述统计量。

(1) 读取数据文件 data07-01, 按 Analyze→Descriptive Statistics→Frequencies 顺序打开主对话框, 选择 age 和 educ 变量进入 Variable(s)框中, 选中 Display frequency tables, 要求显示频数分布表。

(2) 单击 Statistics 按钮, 在相应对话框中, 选 Percentile Values 栏中的 Quartiles。选 Distribution 栏中的 Skewness 和 Kurtosis, 检查数据的正态性。在 Centre Tendency 栏选 Mean(均值)和 Median(中位数), 在 Dispersion 栏选 Std.deviation(标准差)、Minimum(最小值)、Maximum(最大值)和 Range(全距), 单击 Continue 按钮, 返回主对话框。

(3) 单击 Charts 按钮, 打开 Charts 对话框, 选择 Histograms 和 With normal curve。单击 Continue 按钮, 返回主对话框。

(4) 在主对话框中, 单击 OK 按钮, 提交运行。

部分输出结果与分析见表 7-3、表 7-4、图 7-5、图 7-6。

表 7-3 为年龄和受教育年限变量的描述统计量。以年龄为例看输出结果，均值是 45.63，中位数是 41；二者相差较大，说明 age 变量是偏态的。偏度为 0.524，大于 0，说明 age 左偏，有一个较长的右尾，峰度值为-0.786，小于 0，曲线比较平缓；可以对照直方图认识这个变量。对变量 educ 读者可以自己从输出表中认识它。

表 7-3 年龄与受教育年限变量的描述统计量

Statistics		年龄	受教育年限
N	Valid	1514	1510
	Missing	3	7
Mean		45.63	12.88
Median		41.00	12.00
Std. Deviation		17.808	2.984
Skewness		.524	-.168
Std. Error of Skewness		.063	.063
Kurtosis		-.786	.710
Std. Error of Kurtosis		.126	.126
Range		71	20
Minimum		18	0
Maximum		89	20
Percentiles	25	32.00	12.00
	50	41.00	12.00
	75	60.00	15.00

表 7-4 受教育年限变量的频数分布表

受教育年限				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	.2	.1	.1
	3	.5	.3	.5
	4	.5	.3	.8
	5	.6	.4	1.2
	6	12	.8	2.0
	7	25	1.6	3.6
	8	68	4.5	8.1
	9	58	3.7	11.9
	10	73	4.8	16.7
	11	85	5.6	22.3
	12	461	30.4	52.8
	13	130	8.6	61.5
	14	175	11.5	73.0
	15	73	4.8	77.9
	16	194	12.8	90.7
	17	43	2.8	93.6
	18	45	3.0	96.6
	19	22	1.5	98.0
	20	30	2.0	100.0
Missing	NA	1510	98.5	100.0
Total		1517	100.0	

图 7-5、图 7-6 为 age 和 educ 变量带有正态曲线的直方图，从图中可以比较明显地看到数据的分布与正态分布不一致，这与偏度、峰度值的结果一致。age 变量有一个较长的右尾，曲线较平缓。educ 变量右偏，左尾较长，曲线较陡峭。

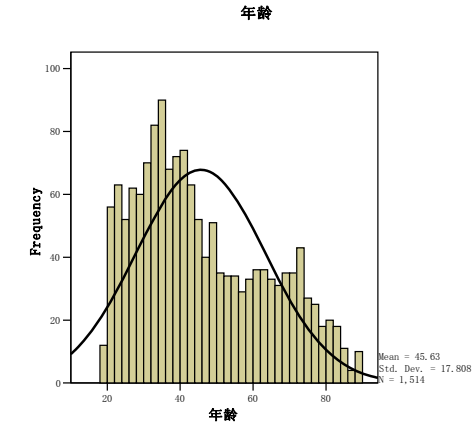


图 7-5 age 变量的直方图

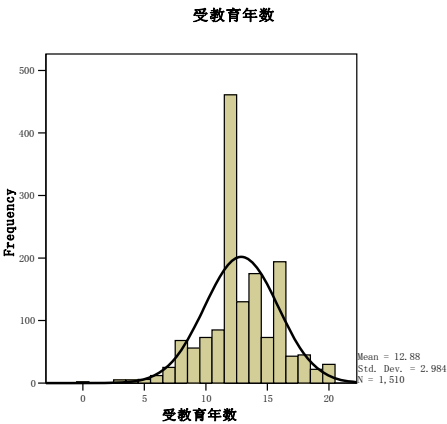


图 7-6 educ 变量直方图

在这里需要说明的是（1）图中正态曲线是根据变量的均值和标准差绘制的，不是标准正态分布。（2）age 和 educ 变量的值都较多，最好先分组，然后再编制频数分布表。

## 7.2 描述统计

描述统计分析过程通过计算均值、算术和、标准差、最大值、最小值、方差、全距

和均数的标准误等统计量对变量进行描述。通过  $z$  分数探明异常观测量。描述统计分析过程适用于尺度变量。至于使用哪些统计量做最终描述,要看其正态性检验的结果。

### 7.2.1 描述统计中的基本概念

1. 均值、中位数和众数都是反映数据集中性特征的统计指标。当数据分布呈均匀分布或正态分布时,均值是反映一组数据平均水平常用的统计指标。当数据分布不对称或有极端值时,中位数是反映数据平均水平的一个较好的统计指标。初步认识一组数据,可以使用众数,它是指一组数据中出现次数最多的那个数。如果中位数与众数相差很大,说明变量值中存在异常值。如果均值和中位数相差太大,说明数据的分布是偏态的。具体的计算公式参见第8章。

2. 四分位数和百分位数都是描述变量值相对位置的统计指标。四分位数是指将一组数据按从小到大的顺序排序后,将其分成四等份,每个等分点上的值即为一个四分位数。百分位数是指将排序后的数据分成100等份,每个等分点上的值即为一个百分位数。较常用的百分位数是第5和第95百分位数。通过计算百分位数,可以了解某个值在集体中的位置。比如我们收集到某个班50名学生的统计考试成绩,用该数据计算第60百分位数是80分,说明该班中有60%的学生成绩都在80分以下,有40%的学生成绩高于80分。四分位数实际就是百分位数中的第25百分位数、第50百分位数、第75百分位数,而且第50百分位数也就是中位数。

3. 全距、方差、标准差和标准误都是描述一组数据离散程度或变异大小的统计指标。具体的计算公式参见第8章。对正态分布数据常将均值和标准差结合在一起描述一组数据的特征。标准误是反映抽样误差大小的统计指标。标准误有均数的标准误和率的标准误,是样本均数(或样本率)的标准差。标准误是由样本均数(或样本率)推断总体均数(或总体率)可靠程度的统计指标。标准误小,说明样本均数(或样本率)与总体均数的差异小,抽样误差小,由样本均数(或样本率)推断总体均数(或总体率)的可靠性程度则高。

4. 偏度和峰度是描述数据分布状况的统计指标。偏度,也称为偏斜度,描述数据分布的偏斜程度和方向。正态分布的偏度值为0。偏度值为正值,分布左偏,右侧有长尾;偏度值是负值,则分布右偏,左侧有长尾。一个经验参考是,如果计算的偏度值在-1到1之间,则表明数据分布近似对称分布。峰度是描述数据分布曲线陡峭平缓程度的统计量。正态分布的峰度值是0。如果峰度值为正,分布曲线是比较陡峭,其峰比标准正态分布的峰高,两端的尾部较长;如果峰度值为负,表明分布曲线是比较平缓的,其峰比标准正态分布的峰低,两端的尾部较短。

### 7.2.2 描述统计分析过程

描述统计分析过程主要计算数据的集中趋势和离中趋势指标。操作和内容如下:

(1) 按 Analyze→Descriptive Statistics→Descriptives 顺序打开对话框,如图7-7所示。



- (2) 在源变量表中选择一个或多个变量作为待分析变量移入 Variable(s)框中。
- (3) 选中 **Save standardized values as variables**，对所选择的每一个变量进行标准化，产生相应的  $Z$  分值，作为新变量保存在数据窗口中。其变量名为相应变量名加前缀  $z$ 。变量标准化的计算公式为

$$Z_i = \frac{x_i - \bar{x}}{s}$$

式中， $x_i$  为变量  $x$  的第  $i$  个观测值， $\bar{x}$  为变量  $x$  的均值， $s$  为变量  $x$  的标准差。

- (4) 单击 **Options** 按钮，展开如图 7-8 所示的选项对话框。在对话框中可以指定其他统计量与输出结果显示的顺序。

基本统计量参见 7.2.1 节中的介绍。

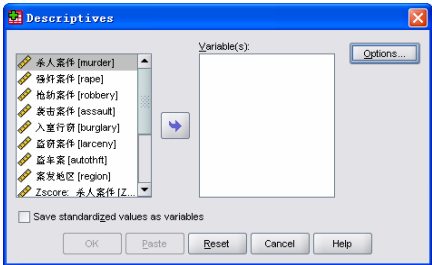


图 7-7 描述统计分析主对话框

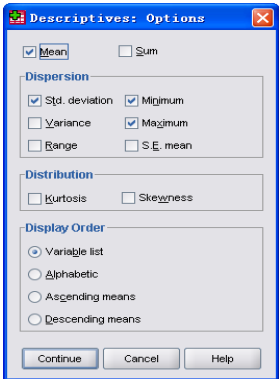


图 7-8 选项对话框

7.2.3 描述统计分析实例

【例 3】数据 data07-02 是对 1985 年美国联邦调查局对 50 个州各种犯罪情况调查的数据。变量：murder、rape、robbery、assault、burglary、larceny、autothft 分别为：谋杀、强奸、抢劫、袭击、入室行窃、盗窃、盗车的案件数。对该数据进行描述统计分析。

- (1) 打开数据文件，按 **Analyze→Descriptive Statistics→Descriptives** 顺序打开如图 7-7 所示的描述统计分析主对话框。
- (2) 将 murder、rape、robbery、assault、burglary、larceny、autothft 变量送入 Variable(s) 栏中。
- (3) 选中 **Save standardized values as variables** 要求计算变量的标准化值，并保存到当前数据文件。
- (4) 单击 **Options** 按钮，打开选项对话框。选中 **Mean**、**Sum**、**Std. deviation**、**Minimum**、**Maximum**、**Range** 要求计算的描述统计量。
- (5) 单击 **Continue** 按钮返回主对话框。单击 **OK** 按钮提交运行。输出结果见表 7-5。

表 7-5 中, 从左至右分别为变量名称、样本量、全距、最小值、最大值、算术和、均数及标准差。最后一行为有效样本量, 本例是 50。

表 7-5 全美各种犯罪数据描述统计量

Descriptive Statistics							
	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation
杀人案件	50	15	1	15	343	6.86	3.848
强奸案件	50	32	4	36	781	15.62	7.348
抢劫案件	50	437	7	443	5076	101.51	91.193
袭击案件	50	272	21	293	6771	135.42	68.170
入室行窃	50	1467	286	1753	46540	930.80	361.050
盗窃案件	50	2856	694	3550	97182	1943.64	709.829
盗车案	50	800	78	878	18393	367.86	199.610
Valid N (listwise)	50						

## 7.3 探索分析

### 7.3.1 探索分析的意义和数据要求

1. 探索分析过程提供对测得数据在分组与不分组的情况下的考查。考查可以有以下两个方面:

#### (1) 检查数据是否有错误

过大或过小的数据均有可能是异常值、影响点或是错误输入的数据。对于这样的数据第一要找出, 第二要分析原因, 第三要决定是否从后续的分析中剔除。因为异常值和影响点往往对分析结果影响较大, 不能真实地反映数据的总体特征。

#### (2) 验证数据分布特征

许多分析方法对数据的分布有要求。检查数据是否为正态分布以便选择分析方法。

另外对若干组数据均值差异性的分析需要根据其方差是否相等, 选择进行检验的计算公式。所以, 分析之前需要首先考查其方差的齐次性等。

### 2. Explore 过程对变量和数据的要求

Explore 过程要求参与分析的变量是等间隔测度的数值型变量。分类变量是数值型或是字符型。箱图中用来标识异常值的变量可以是字符型的也可以是数值型。

### 3. Explore 提供的考查方法

Explore 过程除了输出描述统计量外, 还提供图形可以直观地将异常值、非正常值、缺失数据及数据本身的特点呈现出来。进行假设检验, 为读者选择分析方法提供依据。

(1) 箱图, 如图 7-9 所示。它是对任何分布的数据的整体描述。

① 矩形框是箱图的主体, 上中下三条线分别表示变量值的第 75、50、25 百分位数。

② 中间的纵向直线称触须线。上截止横线是变量值本体最大值, 下截止横线是变



K-S 统计量。显著水平值  $\text{Sig} < 0.05$  时, 拒绝正态分布假设。

#### (4) 方差齐性检验

许多检验要求方差齐性。例如方差分析要求各分组样本的数据来自方差相同的正态总体; 在进行独立样本的 T 检验之前也需要事先确定两组方差是否相同。在进行多个均值组间比较时, 也需要对方差是否相等进行选择。

如果各组方差不等, 可以对数据进行转换来稳定方差或使方差尽可能地相等。

##### ① Spread vs. Level 图

Spread vs. Level 图用来判断各组离散程度是否相同。显示图形的同时, 还输出回归方程斜率以及对数据进行幂转换的幂值。它们之间的关系为: 幂值 =  $1 - \text{回归斜率}$ 。如果没有指定因素变量, 则不生成此图。

② Levene 检验, 该方法最大的好处是对两个样本的数据进行方差齐性检验时, 不强求数据必须服从正态分布。Levene 检验法先计算出各个观测值减组均值的差, 然后再通过这些差值的绝对值进行单因素方差分析。如果显著水平值小于 0.05, 则拒绝各组方差相等的假设。如果选择了数据转换, Levene 检验是根据转换后的数据计算的。

在进行方差齐性检验时, SPSS 提供了四种指标进行判断。四种指标分别是依据均值、依据中位数、依据中位数与调整后的自由度、依据调整的均值所得的各个统计量。

这是因为这几种统计量各有利弊。均值容易受到最大值、最小值以及极端值的影响。后三种都是比较不错的方法, 但它们都是将极端值排除在外, 调整均值则排除了一部分观测测量数据。

③ M 集中趋势最大似然比的稳健估计统计量。它是样本数据均值与中位数统计量的另外一种表现形式。当数据的分布具有较长尾部或者具有极值时, M 估计统计量要比均值以及中位数给出更精确的结果。

M 估计统计量在计算时对所有观测测量加权。权重随观测测量距离分布中心的远近而变, 计算时包括极端值。极端值由于靠外, 因此比位于中心部位的观测测量给予权重较小。M 估计不要求变量值呈正态分布。当数据分布均匀, 并且两尾较长, 或者当数据中存在极端值时, M 可以给出比均值或者中位数更合理的估计。

M 估计方法有 Huber、Andrew、Hampel 和 Tukey。实践说明, 这四种方法都可以很好地取代平均值以及中位数, 其中 Huber 估计方法对于近似正态分布的数据效果最好。

### 7.3.2 探索分析过程

1. 按 Analyze→Descriptive Statistics→Explore, 打开如图 7-11 所示的对话框。

2. 从源变量框中, 选择若干个数值型变量作为因变量送入 Dependent 框中。此时单击 OK 按钮即可获得默认的统计分析, 这其中包括箱图、茎叶图以及基本的描述统计量。默认情况下缺失值将会被排除到分析过程之外。

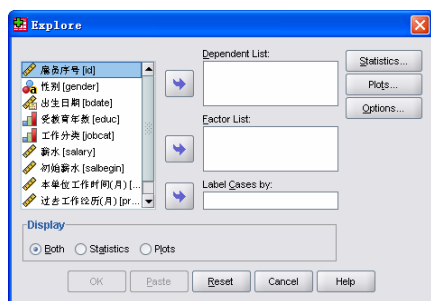


图 7-11 探索分析主对话框

#### 4. 选择标识变量

在源变量表中指定一个变量作为观测量的标识变量，送入 Label Cases by 框中。

#### 5. Display 栏，确定输出项

(1) Both，输出图形以及描述统计量。

(2) Statistics，只输出描述统计量。

(3) Plots，只输出图形。

#### 6. 选择描述统计量

单击 Statistics 按钮，打开如图 7-12 所示的对话框。

(1) Descriptives，可选择的基本描述统计量有：均值、中位数、众数、5%的调整均值、标准误、方差、标准差、最大值、最小值、全距、四分位数、峰度与偏度，及峰度与偏度的标准误。

在 Confidence intervals for mean 框中设置均值的置信区间。在参数框中输入置信水平，选择的范围从 1% 到 99%，常用的数值为 90%、95%、99%。95% 为默认值。

(2) M-estimators，输出集中趋势最大似然比的稳健估计。

(3) Outliers，输出 5 个最大值与最小值，在输出窗口中它们被标明为极端值。

(4) Percentiles，输出第 5、10、25、50、75、90 以及 95 百分位数。

#### 7. 统计图形及其参数的选择，展开 Plots 统计图对话框，见图 7-13。

(1) Boxplots 栏，箱图选项。除 None 是不显示箱图外：

① Factor levels together，因变量按因素变量分组，各组的因变量生成的箱图并列。

② Dependents together，所有因变量在一个图形中生成各组箱图，利于比较。

(2) Descriptive 栏，选择描述图形。Stem-and-leaf，生成茎叶图，这是默认选项；Histogram，生成直方图。

(3) Normality plots with tests，输出正态概率与无趋势概率图。同时输出 K-S 统计量及其 Lilliefors 显著性概率，Shapiro-Wilk 统计量及其显著性概率。

(4) Spread vs level with levene test 栏，输出 Spread vs level 图同时输出回归直线斜率以及方差齐性的 Levene's 检验结果。如果没有指定分组变量，此选项无效。如果选择了

#### 3. 指定分组变量

在源变量框中选择一个或多个分组变量进入 Factor 框中。分组变量可以将数据按该变量中的观测值进行分组分析。如果选择的分组变量不只一个，那么会以分组变量各取值进行组合分组。例如指定分组变量：性别 sex (f、m)、年龄段 age (11、12、13) 为，则按组合分组为：(f,11)、(f,12)、(f,13)、(m,11)、(m,12)、(m,13)，分组对数据进行分析。

Transformed 转换选项，将依据转换后的数据进行计算。

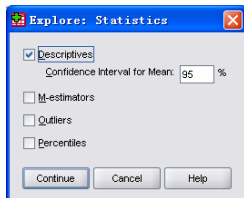


图 7-12 描述统计量对话框

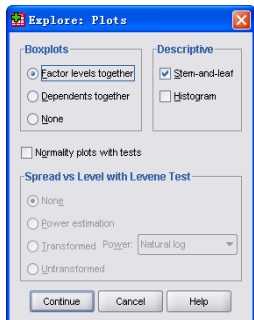


图 7-13 统计图对话框

① None，不产生 Spread vs Level 图，不进行方差齐性的 Levene 检验，是默认选项。

② Power estimation，估计幂值。对每一组数据产生一个中位数的自然对数与四分位数的自然对数的 Spread vs Level 图。同时为了使每组中的数据方差相等对数据进行幂变换。这个结果常常用来确定转换时最合适的幂值。

- Transformed，对原始数据进行转换，由读者在 Power 参数框中指定幂变换使用的幂值。Cube 幂值为 3，Square 幂值为 2，Square root 幂值为 0.5，Logarithm 幂值为 0，Reciprocal of square root 幂值为 -0.5，Reciprocal 幂值为 -1。

- Untransformed，不对数据进行转换。

8. 单击 Options 按钮，展开如图 7-14 所示的对话框。选择分析过程中对缺失值的处理方式。

(1) Exclude cases listwise，剔除带有缺失值的观测量，这是默认选项。

(2) Exclude cases pairwise 成对剔除有缺失值的观测量。

(3) Report values 分组变量中的缺失值将被单独分为一组。显示在频数表中。

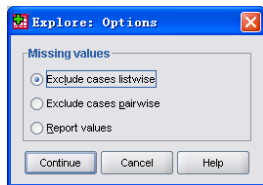


图 7-14 Options 对话框

### 7.3.3 探索分析实例

**【例 4】** data07-03 是 1969~1971 年美国一家银行的 474 名雇员情况的数据，包括变量：salary 当前薪水、educ 受教育年限（年）、prevexp 工作经历（月）、minority 是否是少数民族（0：非少数民族，1：少数民族）、jobcat 工作分类、id 雇员序号等。下面以 salary 当前薪水变量为例，说明探索分析的操作过程及其结果。

1. 选择变量，指定选项。

(1) 打开 data07-03，按 Analyze→Descriptive Statistics→Explore 顺序打开主对话框。

- (2) 选择 salary 变量进入 Dependent List 框，选择 gender 变量进入 Factor List 框，选择 id 变量进入 Label Cases by 框。在 Display 栏中，选择 Both 项。
- (3) 单击 Statistics 按钮，打开对话框。选中 Descriptives、M-estimators、Outliers。
- (4) 单击 Plots 按钮，打开 Plots 对话框。选择 Boxplots 栏中的 Factor levels together；选择 Descriptive 栏内的 Stem-and-leaf，选中 Normality plots with tests；在 Spread vs. Level with Levene Test 栏中选择 Power estimation。单击 Continue 按钮，返回主对话框。
- (5) 在主对话框，单击 OK 按钮，提交运行。

2. 部分输出结果见表 7-6 至表 7-11，图 7-16 至图 7-19。重点进行解释。

表 7-6 是分析变量的描述统计量（注：作者将一个表分成两个表显示。）：因变量 salary；分组变量 gender；Interquartile Range 是四分位数差，其他统计量说明略。女雇员的偏度值为 1.863，峰度值为 4.641，说明变量 salary 的分布不呈正态。

表 7-7 中的 a、b、c、d 分别表示四种 M 估计统计量，它是根据各自的加权常数计算的。加权数在表框的下面。与表 7-6 的均值比较，发现 M 估计值全部要比均值小（Female=\$26031.92，Male=\$41,441.78），且相差较大，据此可初步判定各组数据不是来自正态分布总体。M 估计值与中位数（Female=\$ 24300，Male=\$ 32850）十分接近。

表 7-6 Salary 的描述统计量

Descriptives					Statistic	Std. Error
当前薪水	性别	Mean			\$26,031.92	\$514.258
		95% Confidence Interval for Mean	Lower Bound		\$25,019.29	
			Upper Bound		\$27,045.55	
		5% Trimmed Mean			\$25,248.30	
		Median			\$24,300.00	
		Variance			5.712E7	
		Std. Deviation			\$7,558.021	
		Minimum			\$15,750	
		Maximum			\$58,125	
		Range			\$42,375	
		Interquartile Range			\$7,012	
		Skewness			1.863	.166
		Kurtosis			4.641	.330

Descriptives					Statistic	Std. Error
当前薪水	性别	Mean			\$41,441.78	\$1,213.968
		95% Confidence Interval for Mean	Lower Bound		\$39,051.19	
			Upper Bound		\$43,832.37	
		5% Trimmed Mean			\$39,445.87	
		Median			\$32,850.00	
		Variance			3.602E8	
		Std. Deviation			\$19,499.214	
		Minimum			\$19,650	
		Maximum			\$135,000	
		Range			\$115,350	
		Interquartile Range			\$22,675	
		Skewness			1.639	.152
		Kurtosis			2.780	.302

表 7-8 中 Case number 是观测样品的编号，雇员序号是 id。显示了按性别分组的各组中的 5 个最大值（最高薪水）和 5 个最小值（最低薪水）。

表 7-7 M 估计量

M-Estimators				
	性别	Huber's M-Estimator <sup>a</sup>	Tukey's Biweight <sup>a</sup>	Hampel's M-Estimator <sup>c</sup>
当前薪水	女	\$24,606.10	\$24,015.98	\$24,419.25
	男	\$34,820.15	\$31,779.76	\$34,020.57

- a. The weighting constant is 1.339.  
b. The weighting constant is 4.685.  
c. The weighting constants are 1.700, 3.400, and 8.500.  
d. The weighting constant is 1.340\*pl.

表 7-9 为检验数据是否为正态分布的统计量。自左至右分别为：Kolmogorov-Smirnov 统计量值、自由度、显著性概率，Shapiro-Wilk 检验的统计量、自由度、显著性概率。因于表中 Kolmogorov-Smirnov 下的显著性概率值为 Sig=0.000<0.05，所以拒绝数据呈正态分布的假设。

表 7-10 为方差齐性检验结果。自左至右: Levene 统计量、自由度 1、自由度 2 和显著概率值; 自上至下: 依据均值的结果、依据中位数的结果、依据中位数与调整后的自由度所得的统计量、依据切尾均值所得的各个统计量。依据各种集中趋势统计量所做检验的显著概率值全部低于 0.001, 拒绝方差相等的零假设。即男女薪水方差不具有齐次性。

图 7-15 (a)、(b) 分别为男、女工资水平的茎叶图, 可以推断男雇员工资集中在 25000~39000 之间, 女雇员工资集中在 16000~29000 之间。从图中可以看出男女之间的工资水平可能有较大差异。

图 7-16 (a) 是为男雇员当前薪水的正态 Q-Q 图, 其中的直线是正态分布的标准线, 围绕直线的各点为预测值, 如果观测数据的分布是正态分布, 这些点形成的线应与直线重合。图中大量的点偏离了直线, 因此数据分布不呈正态分布。点组成 V 形曲线, 图 7-16(b) 是无趋势正态 Q-Q 图。也可得出拒绝正态分布的结论。

表7-8 变量的极端值

Extreme Values						
	性别		Case Number		雇员姓名	Value
当前薪水	女	Highest	1	371	371	\$58,125
			2	348	348	\$56,750
			3	468	468	\$55,750
			4	240	240	\$54,375
			5	72	72	\$54,000
	Lowest		1	378	378	\$15,750
			2	338	338	\$15,900
			3	411	411	\$16,200
			4	224	224	\$16,200
			5	90	90	\$16,200
男	Highest	1	29	29	\$135,000	
			2	32	32	\$110,625
			3	18	18	\$103,750
			4	343	343	\$103,500
			5	446	446	\$100,000
	Lowest	1	192	192	\$19,650	
		2	372	372	\$21,300	
		3	258	258	\$21,300	
		4	22	22	\$21,750	
		5	65	65	\$21,900	

表 7-9 正态分布检验结果

Tests of Normality							
	性别	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
当前薪水	女	.146	216	.000	.842	216	.000
	男	.208	258	.000	.813	258	.000

a. Lilliefors Significance Correction

表 7-10 方差齐性检验结果

Test of Homogeneity of Variance					
		Leverage Statistic	df1	df2	Sig.
当前薪水	Based on Mean	119.669	1	472	.000
	Based on Median	51.603	1	472	.000
	Based on Median and with adjusted df	51.603	1	310.594	.000
	Based on trimmed mean	95.446	1	472	.000

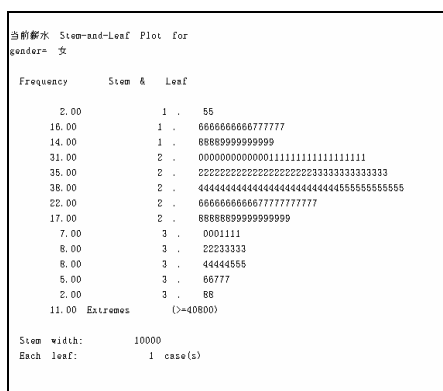
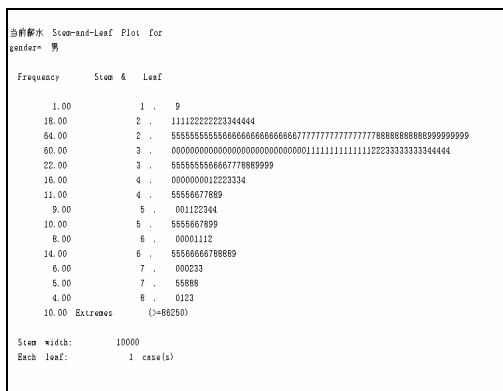
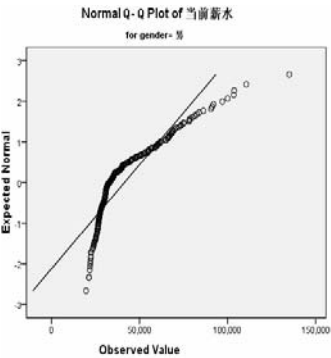
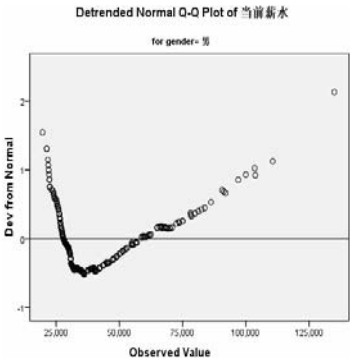


图 7-15 按性别变量分组的 salary 茎叶图





(a)



(b)

图 7-16 当前薪水的 Q-Q 图

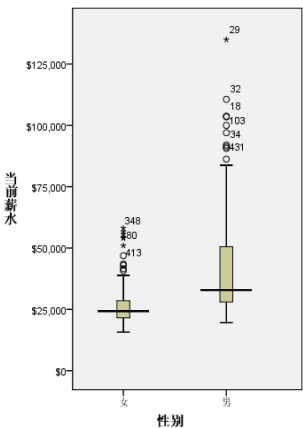


图 7-17 当前薪水的箱图

图 7-17 为性别变量 `gender` 的两个分组的薪水箱图。女雇员当前工资水平的全距较男雇员的小，两组变量中都存在不少异常值。如男雇员中的 29、32、18、103 号的观测值，如女雇员中的 413 号的观测值等。男雇员中 29 号女雇员中的 80、348 都是极值。查看这些观测其他变量值，分析原因，确定后续的分析中是否包括这些观测或按其他变量分组分析。

## 7.4 列联表分析

列联表（又称交叉表）分析过程可以生成二维或多维（分层）频数表，并可以进行分类变量之间的独立性检验。

### 7.4.1 列联表及其独立性卡方检验的思路

#### 1. 列联表的概念

在实际分析中，我们常常将两个分类变量联系起来讨论它们之间是否存在关联，例如收入高低和地区之间是否有关联？收入高低与性别之间是否存在关联？性别与是否喜欢体育锻炼之间是否存在关联等。对于这样问题的研究在统计学中可以使用列联表将两个问题联系起来进行描述。一个变量作为行变量，其值的个数  $r$  即为行数，另一个变量作为列变量，其值的个数  $c$  即为列数，形成  $r \times c$  列联表。最简单的列联表是  $2 \times 2$  的四格表。如性别（男、女）与是否喜欢体育锻炼（喜欢、不喜欢）两个变量的关联性分析就可以通过一个  $2 \times 2$  的四格

表 7-11 百货店与服务满意度交叉列联表

百货店\*服务满意度交叉列联表

		服务满意度					合计
		非常不满意	有些不满意	一般	比较满意	非常满意	
百货店	第一百货店1	25	20	38	30	33	146
	第一百货店2	26	30	34	27	19	136
	第一百货店3	15	20	41	33	29	138
	第一百货店4	27	35	44	22	34	162
合计		93	105	157	112	115	582

表形式进行描述。又如表 7-11 所示的百货店与服务满意度交叉列联表就是一个 4×5 的列联表。表中单元格中的数值是符合行列交叉情况发生的频数。

## 2. 列联表独立性检验基本思路

在统计学中可以通过列联表的独立性检验对两个变量是否存在关联进行分析。该方法的基本思想与假设检验的基本思想是一样的，首先建立一个无效假设，即认为两个事物之间是独立的，没有关联。在假设成立的前提下，建立一个  $\chi^2$  统计量，并计算它发生的概率，根据小概率事件在一次试验中不可能发生的原理，判断建立的无效假设是否成立。若拒绝无效假设，则做出两个事物之间存在关联的判断。因此列联表的独立性检验也称为列联表的  $\chi^2$  检验。 $\chi^2$  统计量的公式为：

$$\chi^2 = \sum \frac{(A - T)^2}{T}$$

式中，A 是实际频数，T 是期望频数

**注意**，使用这个统计量公式进行检验时，要求期望频数大于等于 5。若不满足该条件需要使用精确检验法。

## 7.4.2 列联表分析过程

一个行变量和一个列变量可以形成一个二维列联表，再指定一个分组变量作为控制变量（也称层变量）就形成三维列联表。如果可以指定多个行、列、控制变量，就会形成复杂的多维列联表。列联表的变量可以是数值型、字符型或短字符串变量。

1. 按 Analyze→Descriptive Statistics→Crosstabs 顺序打开如图 7-18 所示的主对话框。
2. 在源变量框中选择一个或多个分类变量送入 Row(s)框，作为列联表中的行变量。
3. 在源变量框中选择一个或多个分类变量送入 Column(s)框，作为列变量。
4. 选择一个层变量进入 Layer 框中。单击 Next 按钮，可指定另外一个控制变量。单击 Previous 按钮可改变前次确定的控制变量。

5. Display clustered bar charts，显示各组中各变量的分类条形图。

6. Suppress tables，只输出统计量，不输出交叉列联表。

7. 单击 Statistics 按钮，打开统计量对话框，如图 7-19 所示。

(1) Chi-square，输出四种卡方检验结果。

① Pearson chi-square test（皮尔逊卡方检验），检验的假设是行、列变量相互独立。当自由度大于 1，单元格频数大于 5 时，检验效果较好，是常用的检验方法。

② Likelihood-ratio chi-square（似然比卡方检验），对数线性模型检验方法之一，也是拟和优度检验方法。当样本量较大时，该统计量服从卡方分布。

③ 2×2 交叉表的 Fisher's exact test（Fisher 精确检验）。当样本数小于 20 或有单元格中的期望频数小于 5 时，使用 Fisher 精确检验是较好的检验方法。

④ 2×2 交叉列联表的 Yete's corrected chi-square test，耶茨校正卡方值为卡方值的校

正值，它总是小于卡方值。

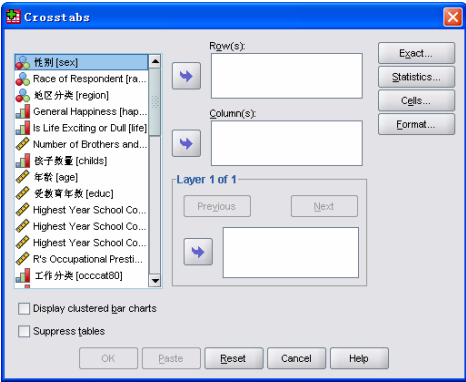


图 7-18 列联表分析主对话框

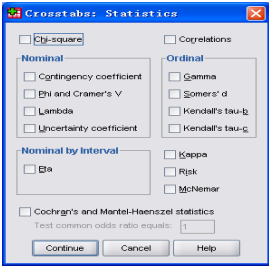


图 7-19 交叉表统计量对话框

(2) Correlations，输出 Pearson 和 Spearman 相关系数。分别表示两变量的线性相关或变量秩之间的关联程度。数值范围为  $-1\sim+1$ 。0 表示无线性关系。符号表示相关方向。如果行、列变量为定序变量，应计算 Spearman 相关系数。

(3) Kappa，输出 Cohen 的 Kappa 系数。用来检验对同一对象两种评估的一致性，它仅适用于具有相同分类值和相同分类数量的变量列联表，如  $2\times 2$  四格表。系数为 1 表示两者完全一致，系数为 0 表示两者没有关联。

(4) Risk，计算 relative risk（相对危险度）和 odd ratio（比数比）。表明事件的发生和某因素之间的关联性。例如，检验心脏病是否与吸烟有关。如果该系数的置信区间包括 1，则认为事件的发生与这个因素没有关联。当某因素发生的可能性非常小时，使用比数比统计量（odd ratio）作为相对危险度的测度。

(5) McNemar，两个二分变量相关性的非参数检验。在“实验前后”的设计中，变化值符合卡方分布。对检验由于实验干扰而产生的变化十分有效。

(6) Nominal 栏，指定名义变量的统计量。

① Contingency coefficient，列联系数是描述两个属性之间关联程度的统计量。根据卡方公式修改而得，公式为

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

式中， $N$  为观测量数， $\chi^2$  为卡方值。其数值在  $0\sim 1$  之间。0 值表示行列变量之间没有关联；其值接近 1，表示行列变量之间有很强的关联。

② Phi and Cramer's V，同列联系数一样， $\phi$  系数和 Cramer V 刻画两个属性间关联程度。根据卡方计算公式修改而得，其值可以达到 1。其计算公式为

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

式中,  $k$  为行、列变量中水平数较小的一个水平数。 $N$  为观测量数、 $\chi^2$  为卡方值。

③ **Lambda**, 当用自变量预测因变量时, 该检验反映预测误差。**Lambda** 系数等于 1, 表明自变量完全预测因变量。**Lambda** 系数等于 0, 表明自变量不能预测因变量。

④ **Uncertainty coefficient**, 不确定性系数。表示用一个变量预测另一个变量的值可能发生的错误程度。其值越接近其上限 1, 表明从第一个变量获得的有关第二个变量的值的信息越多。越接近其下限 0, 表明从第一个变量获得的有关第二个变量的值的信息越少。程序计算对称与不对称两种不确定性系数。

(7) **Ordinal** 栏, 有序变量的统计量。

① **Gamma**, 两个有序变量间关联的对称检验, 该值范围在  $-1 \sim 1$  之间。**Gamma** 的绝对值接近 1, 表明两个变量间高度关联; 接近 0, 表明两个变量间的关联程度很低。对二维列联表, 提供零阶 **Gamma** 值。对三维或高维列联表, 提供条件 **Gamma** 值。

② **Somers'd**, 两个有序变量间关联性的检验, 其数值范围为  $-1 \sim 1$ 。**Somers'd** 的绝对值接近 1 时, 表明两个变量间高度关联。接近 0, 表明低度关联。**Somers'd** 检验是 **Gamma** 的非对称检验的扩展, 两者之间不同仅在于它包含的是未打结自变量成对数据的含量。

③ **Kendall's tau-b**, 秩变量或等级变量关联性的非参数检验, 计算中考虑结的影响。值的范围  $-1 \sim 1$ , 符号表明两者间关系的方向。绝对值表明相关程度, 只有在正方形表格中其值才有可能为 +1 与 -1。

④ **Kendall's tau-c**, 秩变量关联性的非参数检验, 不考虑结的影响。其值的范围是  $+1 \sim -1$ , 符号表明两者间关系的方向, 绝对值表明相关程度。如果交叉表边际频数相等, 那么 **Kendall's tau-b** 和 **Kendall's tau-c** 的值基本一致。

(8) **Nominal by Interval** 栏, 如果一个变量是名义变量, 另一个为定距变量时, 计算 **Eta** 统计量。**Eta** 值的范围在  $0 \sim 1$  之间, 值为 0 表示行列变量间没有关联性, 值越接近 1 关联程度越高。

(9) **Cochran's and Mantel-Haenszel statistics**, 两个二分类变量间独立性检验的统计量。在此框中设置相对风险检验的零假设值, 默认为 1。可以输入一个正数。

8. 单击 **Exact** 按钮, 打开精确检验对话框, 见图 7-20。

除了非参数检验与交叉表检验外, 精确检验提供两种专门针对数据量小或不均衡表的检验方法, 该检验对数据没有要求。检验的方法有 **Fisher** 精确检验和 **Monte Carlo** 法。由于精确检验的计算复杂, 对大样本会耗费大量的计算机资源, 因此在样本量少于 30 时, 这是最好的方法了。

(1) **Asymptotic only**, 显著概率值是基于渐近分布计算

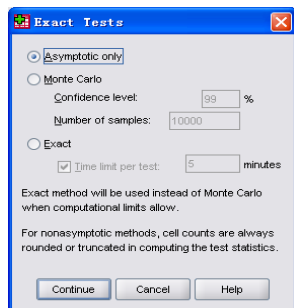


图 7-20 精确检验对话框

的统计量。一般情况下如果其显著水平值小于 0.05, 认为有显著性意义。

(2) Monte Carlo, 该统计量是精确显著水平的无偏估计。Monte Carlo 方法不要求渐近分布的假设, 可获得精确的显著水平值。

① Confidence level 框, 输入 0.01~99.9 置信水平。

② Number of samples 框, 输入 1~1,000,000,000 之间的样本量数值, 用以计算 Monte Carlo 统计量。样本量越大, 显著水平越可靠, 但计算过程耗时也越多。

(3) Exact, 精确计算检验的概率。此值如果小于 0.05, 则认为行、列变量间相互不独立。当期望数有小于 5 的情况时, 适合使用该方法。

选中 Time limit per test 项, 在参数框中输入 1~9,999,999,999 间的值作为进行精确检验的最大运行时间。当计算条件受到限制时, 常使用 Monte Carlo 精确检验法。

9. 在主对话框中, 单击 Cells 按钮, 出现 Crosstabs: Cell Display 对话框, 如图 7-21 所示。

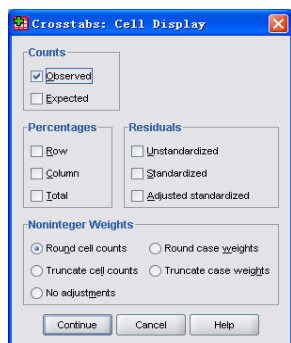


图 7-21 单元格显示对话框

(1) Counts 栏, 指定交叉表中显示的计数选项。

① Observed, 显示实际频数, 这是默认选项。

② Expected, 如果行、列变量在统计意义上相互独立, 显示期望频数 (理论数)。

(2) Percentages 栏, 指定输出的百分比。

① Row, 行百分比, 单元格频数占所在行观测量的百分比。

② Column, 列百分比, 单元格频数占所在列观测量的百分比。

③ Total, 单元格中频数占全部观测量的百分比。

(3) Residuals 栏, 指定要输出的残差。

① Unstandardized, 非标准化残差。单元格中的观测值减期望值。

② Standardized, 均值为 0, 标准差为 1 的标准化残差。残差除以它的标准误, 也称为皮尔逊残差。

③ Adjusted standardized, 调整的标准化残差。

(4) Noninteger Weights 栏, 选择非整数权重处理方法。单元格中的频数一般是整数, 但是如果由带有小数的变量值加权, 单元格的计数值可能出现小数。

① Round cell counts, 照常使用观测量权重, 但是单元格中累积权重需要在计算统计量之前四舍五入。

② Truncate cell counts, 照常使用观测量权重, 但是单元格中累积权重需要在计算统计量之前截取整数部分。

③ Round case weights, 在加权计算之前对权重值四舍五入。

④ Truncate case weights, 在加权之前对权重值截取整数部分。

⑤ No adjustments, 不对单元格数值进行调整。

10. 单击 Format 按钮, 打开 Table Format 对话框, 如图 7-22 所示。确定表格中从左到右频数的排列顺序。

(1) Ascending: 以升序显示变量频数, 这是默认选项;

(2) Descending: 以降序方式显示变量频数。

对长字符型变量可以通过编码满足该过程对数据的要求。

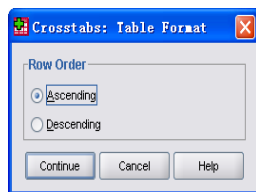


图 7-22 格式对话框

### 7.4.3 列联表分析实例

【例 5】使用 data07-01 中的数据。使用变量: occcat80 工作性质分类、region 地区、childs 每个家庭的孩子数。要求分析各地区工作类型与家庭孩子数之间是否有关联。

(1) 按 Analyze→Descriptive Statistics→Crosstabs 的顺序单击菜单项, 打开主对话框。

(2) 将 childs 作为行变量选入 Row(s)框中, 将 occcat80 作为列变量选入 Column(s)框中。将 region 变量选入 Layer of 框中, 作为层变量。

(3) 单击 Statistics 按钮, 展开 Statistics 对话框, 选中 Chi-square。

(4) 单击 Cells 按钮, 展开 Cell Display 对话框, 在 Counts 栏中选中 Observed。

(5) 打开 Exact 对话框, 选择 Monte Carlo, 在 Number of Samples 中输入样本数量 1517。

(6) 单击 Format 按钮, 展开 Table Format 对话框, 选择 Ascending 项。

(7) 在主对话框中, 单击 OK, 提交执行。

(8) 输出结果表 7-12 至表 7-14, 分析如下。

表 7-12 观测量统计处理摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
孩子数量 * 工作分类 * 地区分类	1414	93.2%	103	6.8%	1517	100.0%

表 7-12 为观测量处理摘要, 其中 N 为个数, Percent 为百分比, Missing 为缺失值。

表 7-13 为不同地区、不同工作性质与不同家庭孩子数量的列联表。Total 为合计。

表 7-14 是独立性的卡方检验结果。

**注意, 由于许多单元格的频数少于 5, 所以应该用 Fisher 精确检验结果得出结论:**

① 东北部地区, Fisher 精确检验 Monte Carlo 显著性概率为 0.117, 大于 0.05, 所以家庭中拥有孩子的数量与工作类型之间没有关系。

② 东南部地区, Fisher 精确检验 Monte Carlo 显著性概率 0.011, 小于 0.05, 所以家庭中拥有孩子的数量与工作类型有关系。

③ 西部地区, Fisher 精确检验显著性概率为 0.072, 大于 0.05, 所以结论与①相同, 即家庭中拥有孩子的数量与工作类型之间没有关联。

表 7-13 各变量之间的多维交叉列联表

孩子数量 ' 工作分类 ' 地区分类 Crosstabulation									
Count		工作分类							Total
地区分类		管理者和专业技术人员	技术人员、销售、行政人员	维修人员	农林渔猎	精密手工制造业	一般操作人员或农民/劳工		
东北部	孩子数量 0	44	57	21	3	19	16		160
	1	27	45	12	3	10	18		115
	2	41	61	23	2	15	27		169
	3	21	43	14	0	8	14		100
	4	12	11	11	0	11	13		58
	5	1	6	3	1	2	6		19
	6	1	2	1	0	0	1		5
	7	3	2	2	1	1	3		12
	8或8以上	0	0	0	0	1	0		1
Total		150	227	87	10	67	98		639
东南部	孩子数量 0	29	31	11	5	8	19		103
	1	22	22	4	4	6	11		69
	2	18	28	22	2	10	20		100
	3	6	11	5	2	10	6		40
	4	3	7	11	0	5	7		33
	5	4	7	0	0	2	5		18
	6	1	1	3	0	0	3		8
	7	0	1	3	0	0	2		6
	8或8以上	0	1	0	0	1	2		4
Total		83	109	59	13	42	75		381
西部	孩子数量 0	37	45	12	5	17	13		129
	1	17	18	6	0	6	12		59
	2	25	29	17	2	10	8		91
	3	11	19	13	3	11	4		61
	4	5	8	2	3	5	4		27
	5	6	0	2	0	2	2		12
	6	2	1	4	0	1	1		9
	7	1	0	1	0	0	0		2
	8或8以上	2	0	1	0	1	0		4
Total		106	120	58	13	53	44		394

表 7-14 卡方检验结果

Chi-Square Tests										
地区分类		Value	df	Asymp. Sig. (2-sided)	Monte Carlo Sig. (2-sided)			Monte Carlo Sig. (1-sided)		
					Sig.	99% Confidence Interval		Sig.	99% Confidence Interval	
						Lower Bound	Upper Bound		Lower Bound	Upper Bound
东北部	Pearson Chi-Square	47.163 <sup>a</sup>	40	.203	.186 <sup>b</sup>	.160	.212			
	Likelihood Ratio	44.483	40	.289	.262 <sup>b</sup>	.233	.291			
	Fisher's Exact Test	48.225			.117 <sup>c</sup>	.095	.138			
	Linear-by-Linear Association	9.514 <sup>c</sup>	1	.002	.003 <sup>b</sup>	.000	.006	.002 <sup>b</sup>	.000	.005
	N of Valid Cases	639								
东南部	Pearson Chi-Square	61.974 <sup>d</sup>	40	.014	.016 <sup>b</sup>	.008	.025			
	Likelihood Ratio	65.957	40	.006	.009 <sup>b</sup>	.003	.016			
	Fisher's Exact Test	55.621			.011 <sup>b</sup>	.004	.018			
	Linear-by-Linear Association	9.398 <sup>e</sup>	1	.002	.003 <sup>b</sup>	.000	.006	.001 <sup>b</sup>	.000	.002
	N of Valid Cases	381								
西部	Pearson Chi-Square	47.883 <sup>f</sup>	40	.183	.191 <sup>b</sup>	.165	.216			
	Likelihood Ratio	52.035	40	.096	.115 <sup>b</sup>	.094	.136			
	Fisher's Exact Test	47.618			.072 <sup>b</sup>	.055	.089			
	Linear-by-Linear Association	.683 <sup>g</sup>	1	.408	.411 <sup>b</sup>	.378	.443	.200 <sup>b</sup>	.174	.227
	N of Valid Cases	394								

a. 有28个(占51.9%)单元格中的期望频数少于 5，最小的期望频数为 0.02。  
b. 根据1517个数据的样本进行 Fisher 检验。  
c. 标准化的卡方值为 3.084。  
d. 有30个(占55.6%)单元格中的期望频数少于5，最小大的期望频数为0.14。  
e. 标准化的卡方值为 3.066。  
f. 有32个(占59.3%)单元格中的期望频数少于5，最小的期望频数为0.07。  
g. 标准化的卡方值为 0.827。

【例 6】小样本的列联表分析实例。

data07-04 为某公司经理收入情况数据。使用变量：sex 性别、earnings 收入高低。分析男、女经理间收入高低是否不同。数据中有 15 个经理，其中男 9 人，女 6 人。由于样本较小，又是 2x2 交叉表，所以使用 Fisher 精确检验的结果。

(1) 操作步骤：读取数据文件后：

① 按 Analyze→Descriptive Statistics→Crosstabs 的顺序打开主对话框。

- ② 将性别变量 sex 选入 Row(s)框中, 将收入变量 earnings 选入 Column(s)框中。
- ③ 单击 Statistics 按钮, 展开 Statistics 对话框, 选中 Chi-square。
- ④ 在主对话框中, 单击 OK 按钮, 提交系统执行。

(2) 输出结果见表 7-15 至表 7-17。

由于样本过小, 所有单元格的期望频数小于 5, 最小的期望频数值为 2.8, Fisher 精确检验计算的双尾概率为  $p=0.041$ , 结论是不同性别之间经理的收入高低差异显著。

表 7-15 观测量处理摘要

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
性别 * 收	15	100.0%	0	.0%	15	100.0%

表 7-16 交叉列联表

性别 * 收入 Crosstabulation				
Count		收入		Total
		低	高	
性别	男性	2	7	9
	女性	5	1	6
Total		7	8	15

表 7-17 卡方检验

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.402 <sup>b</sup>	1	.020		
Continuity Correction <sup>a</sup>	3.225	1	.073		
Likelihood Ratio	5.786	1	.016		
Fisher's Exact Test				.041	.035
Linear-by-Linear Association	5.042	1	.025		
N of Valid Cases	15				

a. Computed only for a 2x2 table

b. 4 cells (100.0%) have expected count less than 5. The minimum expected count is 2.80.

## 7.5 比率分析

常常见到两个尺度类型变量间比值的分析问题。例如, 企业主营业务收入在总收入中的比重; 篮球比赛中三分球得分占总得分的百分比, 财产保险业务保费收入占全部业务保费收入的比例, 汽车功率与车重之比等, 它们都是比率的概念。如果希望计算比率, 并将比率作为一个变量进行描述统计分析, 可以使用比率分析过程。

### 1. 几个基本概念

(1) 平均绝对离差 AAD, 它是各比率值与中位数之差的绝对值之和除以样本量, 即

$$AAD = \frac{\sum |R_i - M|}{N}$$

式中,  $R_i$  是比率变量值,  $M$  是比率变量的中位数,  $N$  是样本量,  $i=1 \sim N$ 。

2. 离散系数 COD, 它是比率变量平均差与中位数的比值, 描述的是比率变量的离散程度。其公式是

$$COD = \frac{|R_i - \bar{R}|}{N \times M}$$

3. 相关价格微分 PRD, 也称为递减指数, 是比率均值与加权比率均值之比。

4. 基于中位数的变异系数 COV, 是对比率变量离散程度的描述, 是比率变量的标准差与中位数的百分比, 其公式为

$$COV = \frac{1}{M} \sqrt{\frac{(R_i - M)^2}{N}}$$

5. 基于均数的变异系数 COV, 与统计学中所讲的变异系数的概念相同, 只是这里



的变量是一个比率变量，它是比率变量的标准差与均数的百分比。

### 7.5.1 比率分析过程

1. 按 Analyze→Descriptive Statistics→Ratio...进入如图 7-23 所示比率分析主对话框。
2. 将计算比率的分子变量送入 Numerator 框。
3. 将计算比率的分母变量送入 Denominator 框。
4. 如果需要进行分组分析，将分组变量送入 Group Variable 框。
5. 单击 Statistics 按钮，进入如图 7-24 所示的统计量对话框，选择要输出的统计量。

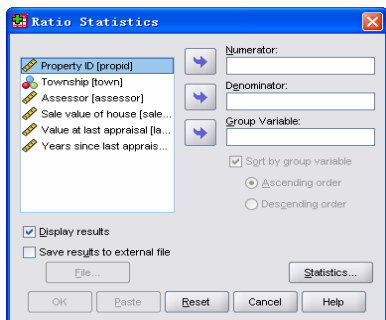


图 7-23 比率分析主对话框

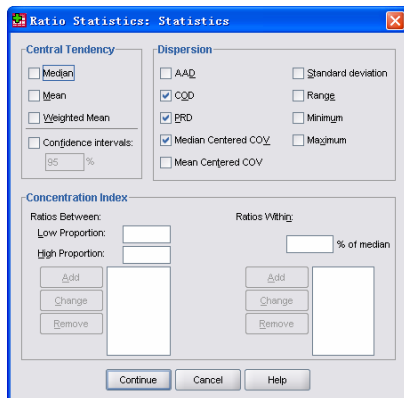


图 7-24 统计量对话框

(1) Central Tendency 栏：中心趋势栏。

- ① Median，输出比率的中位数。
- ② Mean，输出比率的均值。
- ③ Weighted Mean，计算比率的加权均值。该值是用分子的均值除以分母的均值。
- ④ Confidence intervals，计算比率的均数、中位数、加权均值 95% 的置信区间。在框内输入大于等于 0，小于 100 的数值作为置信水平。

(2) Dispersion 栏，输出比率变量离散趋势指标。

- ① AAD，平均绝对离差。
- ② COD，离散系数。
- ③ PRD，相关价格微分。
- ④ Median Centered COV，基于中位数的变异系数。
- ⑤ Mean Centered COV，基于均数的变异系数。
- ⑥ Standard deviation，比率的标准差。
- ⑦ Minimum、Maximum，比率变量的最小值和最大值。
- ⑧ Range，输出比率变量的全距。

(3) **Concentration Index** 集中指数栏：用来测度落在置信区间内的比率百分比集中系数。可以通过两种方式进行计算：

① 在 **Ratios Between** 下面，**Low Proportion** 框内，输入指定区间的下限值，在 **High Proportion** 内输入指定区间的上限值，然后单击 **Add** 按钮，计算落在这个区间的百分比。

② 在 **Ratios Within** 栏定义一个区间，要计算落在距离中位数这个区间内的比率数占比率总数的百分比。可以输入 0 到 100 之间的数值，单击 **Add** 按钮。区间下限为  $(1-0.01 \times \text{该值}) \times \text{中位数}$ 。区间上限为  $(1+0.01 \times \text{该值}) \times \text{中位数}$ 。

6. 在主对话框中选择结果输出方式：

① **Display result**，选择该选项，只在输出窗口显示结果。

② **Save results to external file**，选择该选项，将结果保存为外部文件。

③ **Sorted by group variable**，指定按分组变量输出的结果的升降序。选择 **Ascending order** 或 **Descending order** 按分组变量值的升序（或降序）输出结果。

### 7.5.2 比率分析实例

【例 7】**data07-05** 是美国某州估税员按现有资源价值评估地产价值的假设数据文件。调查数据为过去一年中在该州售出的房产。该数据记录了每处房产在该州的位置（**town**）、估税员自评估以来持续观察地产的时间（**time**）、地产的售价（**saleval**）以及最终估价（**lastval**）。为了帮助政府追踪房产销售状况，合理公正地制定房产税，对估价与售价比进行分析。

(1) 操作步骤

① 按 **Analyze**→**Descriptive Statistics**→**Ratio...**进入比率分析主对话框。

② 选择 **lastval** 变量，将其送入 **Numerator** 框，作为分子变量。

③ 选择 **saleval** 变量，将其送入 **Denominator** 框，作为分母变量。

④ 选择 **town** 变量，将其送入 **group variable** 框，作为分组变量。

⑤ 单击 **Statistics** 按钮，进入 **Statistics** 对话框。在 **Central Tendency** 栏选中 **Median**，在 **Dispersion** 栏取消选中的 **PRD** 和 **Median Centered COV**。

⑥ 在 **Concentration Index** 栏的 **Low proportion** 框内输入 0.8，在 **High Proportion** 框内输入 1.2，单击 **Add** 按钮，将其送入框内；在 **Ration Within** 下的中位数的%框内输入 20，单击 **Add** 按钮，将其送入框内。单击 **Continue** 按钮，返回主对话框。

⑦ 单击 **OK** 按钮，提交运行。结果在表 7-18 和表 7-19 中。

(2) 输出结果及解释

表 7-18 是对样本数据的描述摘要。给出了各地区房产数量和所占百分比。

表 7-19 是房产最终估价与售价的比率（估售比）统计量表。第一列是房产所在位置；第二列是比率的中位数，通过各镇房产估价与售价比率中位数的比较，可以判断哪个位置的房产估售比变化最大。本例中，北部的中位数是 0.963，接近 1，估售比变化最小；

相反，南部房产估售比变化最大；第三列是离散系数 COD，它是描述比率变异大小的指标。数值越大，变异越大。本例中，北部的 COD 是 0.070，最小，说明北部房产的估售比变异最小，而南部的 COD 是 0.199，变异最大；最后两列是集中指数 COC，其中第四列是估售比落在 0.8 和 1.2 之间的百分比，北部是 95.9%，只有 4.1% 房产是需要政府办公室重点关注的。南部该值为 36.1%，有 63.9% 房产需要重点关注。最后一列是估售比落在中位数两侧 20% 区间的占该区所售房产的百分比，其值越大，表明变异越小。

表 7-18 样本数据摘要

Case Processing Summary		
	Count	Percent
房产所在镇的位置	东部	177 17.7%
	中心	187 18.7%
	南部	205 20.5%
	北部	220 22.0%
	西部	211 21.1%
Overall	1000	100.0%
Excluded	0	
Total	1000	

表 7-19 房产最终估价与售价的比率统计量

Ratio Statistics for 房产最终估价 / 房产售价				
Group	Median	Coefficient of Dispersion	Coefficient of Concentration	
			Percent between 0.8 and 1.2 inclusive	Within 20% of Median inclusive
东部	.867	.128	67.2%	78.5%
中心	.904	.118	75.9%	81.8%
南部	.747	.199	36.1%	58.5%
北部	.963	.070	95.9%	95.9%
西部	.816	.118	55.5%	84.8%
Overall	.873	.141	66.3%	75.7%

7.6 P-P图和Q-Q图

P-P 图（probability-probability plot）和 Q-Q 图(quantile-quantile plot) 都是根据累计分布函数理论计算的，使用它们可以进行数据是何种分布的检验，常用于检验数据是否服从正态分布。如果图形中所有点都聚集在直线上，则说明变量分布服从于所要检验的分布。

7.6.1 P-P图和Q-Q图分析过程

- 1. 按 Analyze→Descriptive Statistics→P-P Plots（或 Q-Q Plots）顺序，打开图 7-25P-P（或 Q-Q）图主对话框。Q-Q 图与 P-P 图的界面是一样的。这里以 P-P 图为例进行介绍。
- 2. 将一个或多个被检验的数值型变量送入 Variables 框，对每个变量生成 P-P 概率图。

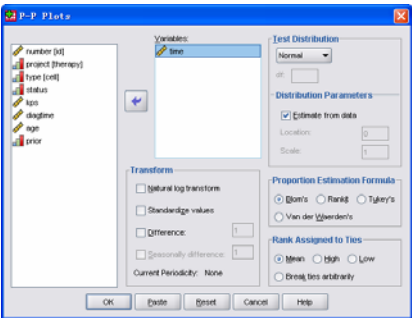


图 7-25 P-P 图对话框

- 3. Test Distribution 指定检验的概率分布。提供 13 种概率分布：Beta 贝塔分布、Chi-square 卡方分布、Exponential 指数分布、Gamma 伽玛分布、Half-normal 半正态分布、Laplace 拉普拉斯分布、Logistic 逻辑斯谛分布、Lognormal 对数正态分布、Normal 正态分布、Pareto 帕雷托分布、Student’ t 即 T 分布、Weibull 威布尔分布、Uniform 均匀分布。
- 如果选择了 T 分布，还需要在 df 自由度参数框中确定自由度。

- 4. 在 Distribution Parameters 栏中选择分布参数，选中 Estimate from data 系统自动从

数据中推算分布的参数, 否则要在参数框中自行指定。选择的分布不同参数框也不同。

5. 在 Transform 栏内选择变量转换方式。

(1) Natural log transform 自然对数转换, 将原变量值转换成以  $e$  为底的自然数值。

(2) Standardize values 标准化转换, 将原变量转换成平均值为 0 和方差为 1 的样本。

(3) Difference 差分转换, 通过计算变量中连续两个数据之差来转换原有变量。输入一个正整数确定差分度。

(4) Seasonally difference 季节差分转换, 计算时间序列中两个恒定间距的数据差, 用来转换原有时间序列数据, 数据间隔的大小是根据当前所选择的周期而定。输入一个正整数以确定差分度。为了计算季节差分, 必须确定含有周期因素的数据变量, 例如一年中的月份。

(5) Current Periodicity 当前周期, 用来指明计算时间序列的季节差分。如果当前周期为 0, 不能计算季节差。在 Data 菜单中 Define Dates 可以建立具有周期性的变量。根据已存在的时间序列使用 Transform 菜单中的 Create Time Series 项, 可以建立新的时间序列变量。

6. Proportion Estimation Formula 比率估计公式栏, 每次只能选择其中一项, 栏内所列公式中,  $n$  是观测量数目,  $r$  是从 1 至  $n$  的秩次。

7. Rank Assigned to Ties 指定结的顺序栏, 一个变量中多个相同的值构成结, 在本栏中可以选用以下不同的方式解决结点处观测量的秩。

(1) Mean 平均秩, 用打结观测值的平均秩来作为它们的秩值。

(2) High 最高秩, 用打结观测值中最高的秩来作为它们的秩值。

(3) Low 最低秩, 用打结观测值中最低的秩来作为它们的秩值。

(4) Break ties arbitrarily 任意拆结, 绘制每个结点处的观测量, 忽视权重的影响。

## 7.6.2 P-P图和Q-Q图分析实例

【例8】打开数据文件 data07-06, 检验肺癌患者生存时间变量 time 是否服从 Weibull 分布。

操作过程: 按 Analyze→Descriptive Statistics→P-P Plots 顺序, 打开 P-P Plots (或 Q-Q Plots) 概率图主对话框。在左侧变量框中选中 time 变量, 送入 Variables 框中。在 Test Distribution 栏内, 单击“下拉”按钮, 选中 Weibull。其他为默认。单击 OK 按钮, 提交运行。

操作结果见表 7-20~表 7-22 和图 7-26。

表 7-20 是模型描述表。主要描述所做 P-P 图的变量名是时间, 未作变量转换, 检验的分布是 Weibull 分布, 估计参数是尺度 (scale) 参数和形状 (shape) 参数。

表 7-21 是样本数据统计摘要。该数据共有 137 个时间序列。无缺失值。

表 7-22 是 Weibull 分布参数估计。该分布尺度参数是 110.127, 形状参数是 0.937。

图 7-26(a)为肺癌生存时间的 Weibull 分布 P-P 概率图。从该图可以看到，各点都在直线上，因此可以得出该数据的分布呈 Weibull 分布。

图 7-26(b)为肺癌生存时间的无趋势 Weibull 分布 P-P 概率图。该图各点是无规则的，表明是随机的。

表 7-20 模型描述

Model Description		
Model Name	MOD_1	
Series or Sequence	1	时间
Transformation	None	
Non-Seasonal Differencing		0
Seasonal Differencing		0
Length of Seasonal Period	No periodicity	
Standardization	Not applied	
Distribution	Type	Weibull
	Scale	estimated
	Shape	estimated
Fractional Rank Estimation Method	Blom's	
Rank Assigned to Ties	Mean rank of tied values	

Applying the model specifications from MOD\_4

表 7-21 样本数据统计摘要

Case Processing Summary		
Series or Sequence Length		时间 137
Number of Missing Values in the Plot	User-Missing	0
	System-Missing	0

The cases are unweighted.

表 7-22 分布参数估计

Estimated Distribution Parameters		
Weibull Distribution	Scale	时间 110.127
	Shape	.937

The cases are unweighted.

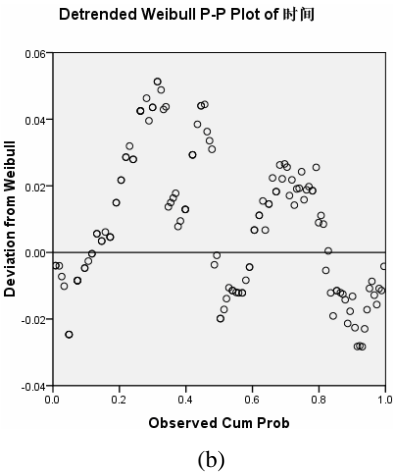
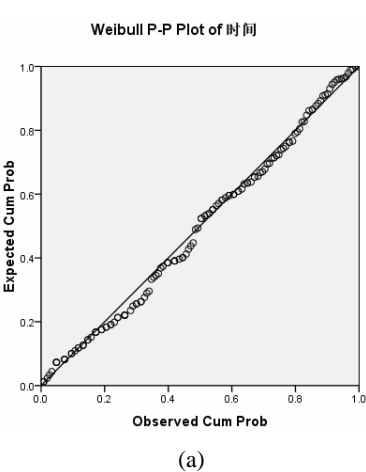


图 7-26 肺癌生存时间的 Weibull 分布 P-P 概率图和无趋势 P-P 概率图

【例 9】data07-07 是 200 例正常人血铅含量数据，用 P-P 概率图分析过程检验 pb 变量是否服从正态分布。

操作步骤：按 Analyze→Descriptive Statistics→P-P Plots 顺序单击菜单项，打开 P-P Plots（或 Q-Q Plots）概率图主对话框。将 pb 变量选入 Variables 框内，在 Test Distribution 框中选择 Normal，其他选项均为默认值，单击 OK 按钮，提交系统运行。程序运行结果参见图 7-27。从图中看到，各点的分布没有完全在直线上，因此得出分布不呈正态分布。

现将 pb 变量的数据转换成自然对数数据，检验转换后的 pb 变量是否服从正态分布。

操作步骤：操作与前例相同。只是在 Transform 栏中选 Natural log transform 选项，结果见图 7-28。

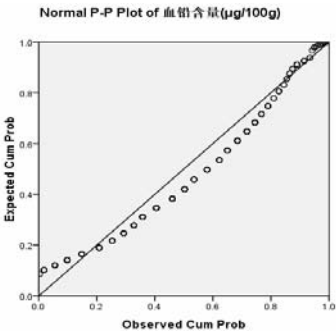


图 7-27 检验 pb 变量正态性的 P-P 图

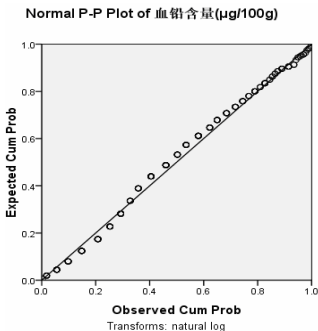


图 7-28 检验转换后 pb 变量正态性的 P-P 图

从图 7-28 可以看出，所有点都在直线上，因此可以得出转换后的数据分布是近似正态的。结论是对非正态数据经过转换呈正态分布，这时就可以用转换后的数据使用正态性假定的方法进行了。

【例 10】 data07-08 是某市 150 名 3 岁女童身高数据，使用 Q-Q 概率图分析过程，检验身高数据的分布是否是正态分布的。

操作过程：按 Analyze→Descriptive Statistics→P-P Plots 顺序，打开 Q-Q Plots 概率图主对话框。将变量 height 选入 Variables 框内，在 Test Distribution 框中选择 Normal 项，其他选项均为默认值，单击 OK 按钮，提交系统运行。程序运行结果参见图 7-29、图 7-30。

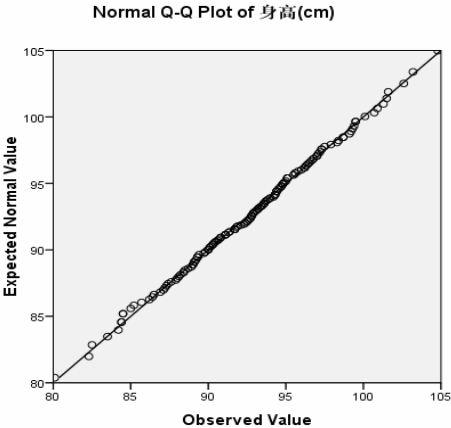


图 7-29 3 岁女孩身高的 Q-Q 概率图

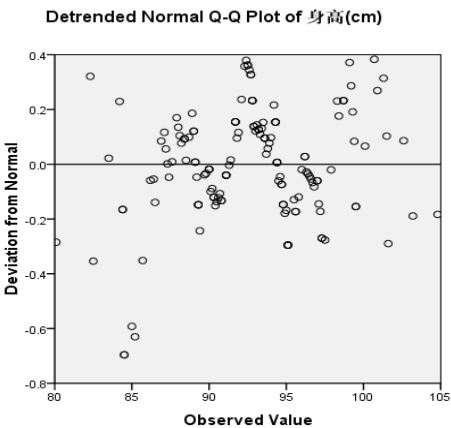


图 7-30 3 岁女孩身高的无趋势 Q-Q 概率图

从图 7-29 可以看到，所有点都在直线上，因此可以得出 150 名 3 岁女童的身高数据分布呈正态分布。图 7-30 的各点是无规则的，表明是随机的。

## 习 题 7

1. 对二维交叉表中两个变量间是否独立的检验，SPSS 提供了几种方法？各适合什么条件？单元格频数小于 5 时，应该考虑用什么方法检验？
2. 正态分布的变量用哪些描述统计量描述？哪些统计量描述该变量值的集中趋势？哪些统计量描述其离中趋势？
3. 如果变量的数据分布不是正态的，可以用哪些统计量描述？
4. 对数据明显为非正态分布的变量能用均值描述其平均水平吗？如果不能，使用什么指标描述比较合适？
5. 打开 data07-09 数据，使用交叉表分析收入类型（`inccat`）与变量订阅报纸（`news`）之间的关系。
6. 使用 data07-09 数据，利用频数表简单说明家庭收入（`income`）数据的分布情况。
7. 使用 data07-10 数据，利用探索过程分析不同质量等级（标准、高级）与合金形成温度是否有关。

## 第 8 章 均值比较与检验

### 8.1 均值比较与均值比较的检验

#### 8.1.1 均值比较的概念

统计分析常常采取抽样研究的方法。即从总体中随机抽取一定数量的样本进行研究来推论总体的特性。由于总体中的个体间存在差异,即使严格遵守随机抽样原则也会由于多抽到一些数值较大的或多抽到一些数值较小的个体致使样本统计量与总体参数之间有所不同。又由于实验者测量技术的差别或测量仪器精确程度的差别等也会造成一定的偏差,使样本统计量与总体参数之间存在差异。由此可以得出这样的认识:均值不相等的两个样本不一定来自均值不同的总体。能否用样本均数估计总体均数,两个变量均值接近的样本是否来自均值相同的总体?换句话说,两个样本中某变量均值不同,其差异是否具有统计意义,能否说明总体差异?这是各种研究工作中经常提出的问题。这就要进行均值比较。

#### 8.1.2 均值比较与检验的过程

SPSS 提供以下计算变量描述统计量的过程和对均值进行检验的过程。

##### 1. MEANS 过程的功能与术语

(1) MEANS 过程计算指定变量的综合描述统计量。当观测量按一个分类变量分组时,MEANS 过程可以进行分组计算。例如要计算工作人员上班路程的平均千米数,SEX 变量把工作人员按性别分为女人、男人两组,MEANS 过程可以分别计算男人、女人上班路程的千米数。用于形成分组的变量应该是其值数量少且能明确表明其特征的变量。可以是标称变量,例如性别、民族、信仰等;也可以是顺序变量,即其值表明等级的,例如年级、职称等。

使用 MEANS 过程求若干组的描述统计量,目的在于比较。因此必须使用分类变量,根据分类变量的值对因变量分组求均值。这是与 Descriptive 过程不同之处。

##### (2) MEANS 过程中使用的术语

① 水平数:指分类变量的值的个数。例如性别变量有 2 个值,称为有两个水平。

② 单元 (CELL):指因变量按分类变量值所分的组。例如可以按性别将因变量的值分为两组。如果还有一个分类变量年龄,共有 10、11、12 三个值,可以将因变量分为



3 组。每组因变量的值称为一个单元。MEANS 过程对每个单元的因变量值,求各种描述统计量。

③ 水平组合:如果有两个分类变量,例如性别(男、女)和年龄(10 岁、11 岁、12 岁)。按它们的水平组合将会分因变量为 6 个单元,即男性 10 岁、男性 11 岁、男性 12 岁、女性 10 岁、女性 11 岁、女性 12 岁。

## 2. T test 过程

T test 过程是对样本进行 T 检验的过程。按不同的比较方式分为 3 个功能:

### (1) 单一样本 T 检验

检验单个变量的均值是否与给定的常数之间存在差异。样本均数与总体均数之间的差异显著性检验属于单一样本 T 检验。例如方便面饼标准重量为 80 克,可以看作总体均数。从生产线上任意抽取 100 个面饼,研究其平均重量与标准重量之间差异是否显著的问题就属于单一样本 T 检验。

### (2) 两个独立样本的 T 检验

两个独立样本的 T 检验用于检验两个不相关的样本来自具有相同均值的总体。例如想知道购买某产品的顾客与不购买该产品的顾客平均收入是否相同,可以使用对两个独立样本进行 T 检验的功能。必须注意,使用这种检验的条件是必须具有来自两个不相关组的观测量,其均值必须是对在两组中相同的变量的测度。

如果分组样本彼此不独立,例如测度的是工人在技术培训前后某项技能的成绩,要求比较培训前后成绩均值是否有显著性差异,应该使用配对 T 检验的功能(Paired Sample T test)。如果分组不只两个,应该使用 One-Way ANOVA 过程进行单变量方差分析。如果试图比较的变量明显不是正态分布的,则应该考虑使用一种非参数检验过程(Nonparametric test)。如果读者想比较的变量是分类变量,应该使用 Crosstabs 功能。

### (3) 配对样本 T 检验

配对样本 T 检验(Paired Sample T test)用于检验两个相关的样本是否来自具有相同均值的总体。这种相关的或配对的样本常常来自这样的实验结果,在实验中被观测对象在实验前后均被观测。例如想要知道技术培训以后是否提高了工作效率,可以在培训前后测试完成一道工序的时间。在构成数据文件时,一个参与测试的工人的培训前后完成一道工序的时间形成一个观测量,两个变量可以命名为 BEFORE 和 AFTER。配对分析的测度也不是必须来自同一个观测对象,一对可以两者组合而成,例如一对夫妻,或者是根据实验前学习成绩和智商均相同的两个孩子作为一对。这样的若干对孩子分为两组,分别用不同教学方法进行教学,一段时间后,比较参与实验的两组学生平均成绩差异是否具有统计意义。在动物实验中常常把同一窝出生的体重、性别相同或最相近的小鼠配成实验的一对。

## 3. One-Way ANOVA 过程

一元方差分析用于检验几个(三个或三个以上)独立的组,是否来自均值相同的总

体。例如可以检验三个减肥训练计划，体重下降的效果（均值）是否相同。同时想看看哪一种训练计划效果最好，或者三个训练计划彼此之间哪两个之间的差异最显著，应该使用 One-Way ANOVA 过程。如果按性别、体重级别对肥胖患者再进行分组，进行三个训练计划的实验。不但想知道哪一种训练计划对降体重下降最有效，而且想知道同一种训练方法对不同性别是否具有不同的效果，或者想去除每天的进食量对训练效果的影响，应该选择 ANOVA Models 子菜单中的各个功能进行多元方差分析或协方差分析。

如果分析变量明显是非正态分布的，应该选择非参数检验过程，非参数检验的内容请见第 12 章。One-Way ANOVA 过程以及多因素方差分析的内容请见第 9 章。

## 8.2 MEANS过程

MEANS 过程的基本功能是分组计算，比较指定变量的描述统计量，包括均值、标准差、总和、观测量数、方差等一系列单变量描述统计量，还可以给出方差分析表和线性检验结果。

使用系统默认值即可按指定分组给出指定变量的均值、标准差、观测量数等基本描述统计量。选项可以给出其他更加丰富的描述统计量。

### 8.2.1 MEANS过程中的统计量

如果变量为  $x$ ， $x_i$  为变量  $x$  的第  $i$  个值，共有非缺失观测量数为  $n$ （或  $N$ ）；如果定义了加权变量  $w$ ，则  $w_i$  为第  $i$  个变量值对应的权重值，可以选择的统计量关键字及含义如下。

1. Sum 总和、加权和公式分别为

$$\text{Sum} = \sum_{i=1}^n x_i \quad \text{Sum} = \sum_{i=1}^n x_i w_i$$

2. Number of Cases 观测量数，如果定义了加权变量为  $w$ ，公式则为

$$N = \sum_{i=1}^n w_i$$

否则所有  $w_i = 1$ ，则  $N=n$ 。

3. Mean 算术平均值，正态分布变量的集中趋势统计量，公式为

$$\text{Mean} = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i}$$

4. Median 中位数，当变量值按大小排序， $N$  为奇数，Median 是中值； $N$  为偶数，Median 是两个中值之平均值。

5. Grouped Median 分组中位数，每组变量值按大小排序， $N$  为奇数，Median 是中值；

$N$  为偶数, Median 是两个中值之平均值。

6. Variance 方差, 正态分布变量的离散趋势统计量, 公式为

$$\text{Variance} = \frac{\sum_{i=1}^n w_i (x_i - \bar{x})^2}{\sum_{i=1}^n w_i - 1}$$

7. Standard Deviation 标准差公式:  $S = \sqrt{\text{Variance}}$

8. Standard Error of Mean 均值的标准误, 公式为

$$\text{Stderr} = \frac{S}{\sqrt{N}}$$

9. Minimum 最小值, 要求  $N \geq 1$ 。

10. Maximum 最大值, 要求  $N \geq 1$ 。

11. Range 范围, 或称全距:  $\text{Range} = \text{Minimum} - \text{Maximum}$ 。

12. First 按分组变量分组的该组的第一个变量值。

13. Last 按分组变量分组, 该组最后一个变量值。

14. Kurtosis 峰度, 是正态性检验统计量之一。其值为负, 分布曲线峰值高出正态分布曲线峰值; 其值为正, 分布曲线比较平坦。公式如下, 要求  $N \geq 3$ ,  $S > 0$ 。

$$\text{Kurtosis} = \frac{N^2 - 2N + 3}{(N-1)(N-2)(N-3)} \cdot \frac{\sum (x_i - \bar{x})^4}{S^4} - \frac{3(2N-3)}{N(N-1)(N-2)(N-3)} \cdot \frac{\left[ \sum (x_i - \bar{x})^2 \right]^2}{S^4}$$

15. Standard Error of Kurtosis 峰度的标准误。

16. Skewness 偏度, 是正态性检验统计量之一。其值为正, 分布曲线相对于正态曲线左偏, 右尾较长; 其值为负, 分布曲线右偏, 左尾较长。公式如下: 要求  $N \geq 2$ ,  $S > 0$ 。

$$\text{Skewness} = \frac{N}{(N-1)(N-2)} \cdot \frac{\sum (x_i - \bar{x})^3}{S^3}$$

17. Standard Error of Skewness 偏度的标准误。

18. Percent of Total Sum 每组总和占整个观测量总和的百分比。

19. Percent of Total N 每组中观测量总数  $N$  占总观测量数的百分比。

20. Geometric Mean 几何均数, 主要用于变量值之间呈倍数关系的偏态分布, 公式为

$$G = \lg^{-1} \left( \frac{\sum \lg x_i}{N} \right)$$

21. Harmonic Mean 调和均数, 主要用于求平均率、平均速度或平均存活时间等。公式为

$$H = \frac{N}{\sum \frac{1}{x_i}}$$

## 8.2.2 MEANS过程操作

### 1. 建立数据文件

数据文件中要求至少有一个连续变量、一个分类变量（离散变量）。对连续变量求其基本描述统计量，分类变量用来分组。

2. 按 Analyze→Compare Means→Means 顺序单击鼠标左键，打开 Means 均值过程对话框，如图 8-1 所示。

### 3. 选择因变量

在左面的变量表中选择要分析的变量作为因变量，送入 Dependent List 因变量列表框中。因变量可以选择一个，也可以选择多个。

### 4. 自变量的选择及层控制

选择分组变量（也称自变量），对因变量将按自变量分组计算基本描述统计量。选择的若干自变量可以放在第一层，也可以放在不同层。

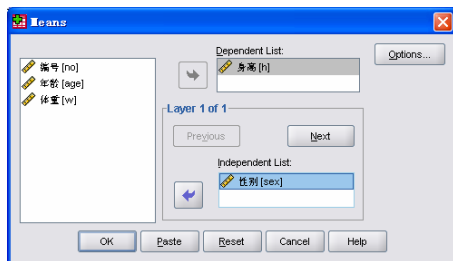


图 8-1 均值比较主对话框

(1) 两个分类变量均放在第一层的操作是：

① 首先在源变量表中选择一个分类变量，送入 Independent List 表中。此时层控制显示 Layer 1 of 1，表示变量被送入第一层。建立了一个控制层。

② 在源变量表中选择第二个变量，送入 Independent List 框中。此时层控制仍显示 Layer 1 of 1，表示变量被送入第一层，共建立了一个控制层。该层有两个自变量。

例如，第一控制层的两个自变量分别有  $n_1$ 、 $n_2$  个水平，则程序运行结果，分别给出两个变量各水平的因变量的统计量。即按第一自变量分  $n_1$  组给出因变量的描述统计量；按第二个自变量分  $n_2$  组给出因变量的描述统计量。

(2) 两个分类变量分别放在两层中的操作是：

① 在变量表中选择一个分类变量，送入 Independent List 框中。建立了一个控制层。

② 单击 Next 按钮，使层控制显示 “Layer 2 of 2”，表明可以建立第二层了。

③ 在变量表中选择第二个分类变量，将其送入第二层，显示在 Independent List 框中作为第二层的分类变量。此时 Previous、Next 两个按钮均加亮。表示既可以单击 Previous 向前回到第一层，又可以单击 Next 按钮，去建立第三层。

如果两个分类变量的水平数分别为  $n_1$ 、 $n_2$ ，并分别控制第一层和第二层，那么会将因变量分为  $n_1 \times n_2$  组，每个组合叫做一个单元（Cell），按单元给出因变量的统计量。

综上所述，单元数的计算是同层变量的水平数相加，不同层的变量水平数相乘。

### 5. Means 过程的选项

在主对话框中用鼠标单击 Options 按钮，展开选项对话框，如图 8-2 所示。

(1) Statistics 栏, 选择统计量。

左面 Statistics 栏内列出了可以计算的各组描述统计量, 选择后单击向右箭头按钮将选定的统计量移至右面 CELL 的矩形框中。可以选择的统计量关键字及含义如下:

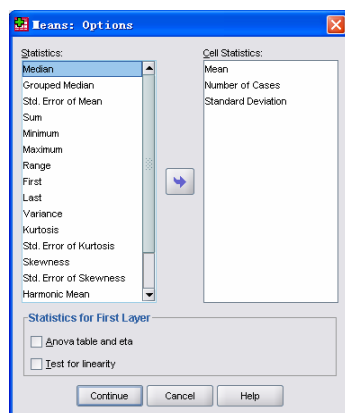


图 8-2 Means 过程选项对话框

Sum (总和)、Number of Cases (观测量数目)、Mean (算术平均值)、Median (中位数)、Grouped Median (分组中位数)、Variance (方差)、Standard Deviation (标准差)、Standard Error of Mean (均值的标准误)、Minimum (最小值)、Maximum (最大值)、Range (范围)、First (按分组变量分组的该组的第一个变量值)、Last (按分组变量分组该组最后一个变量值)、Kurtosis (峰度)、Standard Error of Kurtosis (峰度的标准误)、Skewness (偏度)、Standard Error of Skewness (偏度的标准误)、Percent of Total Sum (每组总和占总和的百分比)、Percent of Total N (每组中观测量总数  $N$  占总观测量数的百分比)、Geometric Mean (几何均值)、Harmonic Mean (调和均值)。

(2) Statistics for First Layer 栏, 指定只对第一层每个控制变量进行的分析。

① Anova table and eta, 方差分析表和 eta 统计量  $\eta$ 、eta square 统计量  $\eta^2$ 。方差分析检验的零假设是, 第一层控制变量各水平上的因变量均值都相等。 $\eta$  统计量表明因变量和自变量之间联系的强度。 $\eta^2$  是因变量中不同组中差异所解释的方差比, 是组间平方和与总平方和之比。

② Tests for linearity, 产生平方和、自由度、均方、F 检验的  $F$  值、 $R$  和  $R^2$  等统计量。但在分类自变量是字符型时, 不计算有关线性度的统计量。 $R$  和  $R^2$  是线性拟合的良好度的统计量, 只有在控制变量有基本的数量级 (例如控制变量表示年龄或药物剂量, 不能是颜色或信仰等), 且自变量有三个水平以上时才计算。其假设的前提是因变量均值是第一层控制变量的线性函数。

### 8.2.3 分析实例

【例 1】使用 data08-01, 27 名男女学生身高数据。数据文件中的变量顺序是: no 编号、sex 性别、age 年龄、h 身高、w 体重。要求按年龄分组比较身高均值; 按性别分组比较身高均值。分析不同年龄和性别的学生身高均值。

(1) 对于不同年龄、不同性别学生的身高的分析, 要把两个分类变量均放在第一层, 操作如下:

① 按 Analyze→Compare Means→Means 顺序单击鼠标, 打开 Means 过程对话框, 如图 8-1 所示。

- ② 在源变量表中选择变量  $h$  作为因变量，送入 **Dependent List** 因变量列表框中。
- ③ 在源变量表中选择分类变量  $sex$ ，送入 **Independent List** 表中。再在源变量表中选择变量  $age$ ，也送入 **Independent List** 框中。建立了一个控制层，该层有两个分类变量。
- ④ 运行的程序语句如下：

MEANS

TABLES=h BY sex age

/CELLS MEAN COUNT STDDEV.

MEANS 语句调用 Means 过程。TABLES 的等号后面是分析变量  $h$ ，BY 后面指定分类变量为  $sex$  和  $age$ 。

CELLS 子命令指定要求计算的描述统计量为均值、观测值总数和标准差。

- ⑤ 运行结果见表 8-1 和表 8-2。

表 8-1 观测值处理汇总表

Case Processing Summary						
	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
身高 * 性别	27	100.0%	0	.0%	27	100.0%
身高 * 年龄	27	100.0%	0	.0%	27	100.0%

表 8-2 基本描述统计量

身高 * 性别				身高 * 年龄			
身高				身高			
性别	Mean	N	Std. Deviation	年龄	Mean	N	Std. Deviation
女	1.5154	13	.06253	10	1.4488	8	.02167
男	1.5357	14	.07623	11	1.5209	11	.03910
Total	1.5259	27	.06941	12	1.6129	7	.01704
				13	1.5900	1	.
				Total	1.5259	27	.06941

(a)

(b)

表 8-1 给出的是汇总信息。在第一控制层只给出两个分类变量：性别、年龄。因此进行的是（身高\*性别）按性别分组的身高均值比较和（身高\*年龄）按年龄分组的身高均值比较。该表 **Included** 栏给出的是，参与每个分析的观测值数  $N$  均为 27 和占总观测值数的百分比 **Percent** 均为 100%。**Excluded** 栏给出每个分析中剔除的观测值数  $N$  均为 0，占总观测值数的百分比均为 0%。**Total** 栏给出观测值总数  $N$  和总百分比。

在 SPSS 统计分析过程执行的输出结果中，均给出这样的汇总表。在以后的章节中，如无特殊需要，不再进行解释或说明或不再列出。

表 8-2 给出的是按性别和按年龄分组的分析结果。由于在定义系统参数时，要求输出显示变量标签和值标签，因此在表格中显示的分类变量不是原变量名和变量值，而是变量标签和值标签。

表 8-2(a)分析变量是身高  $h$ ，分类变量是性别  $sex$ 。可以看出，女生 13 人平均身高 1.5154，标准差为 0.06253；男生 14 人平均身高 1.5357，标准差为 0.07623。27 个学生总平均身高 1.5259，标准差为 0.06941。

表 8-2(b)按年龄分组的结果是 10 岁 8 人平均身高 1.4488，标准差为 0.02167；11 岁 11 人平均身高 1.5209，标准差为 0.0391；12 岁 7 人平均身高 1.6129，标准差为 0.01704；13 岁 1 人平均身高 1.59，不能计算标准差，因此该项为缺失值。

(2) 另一种分析。发育阶段相同年龄的男孩和女孩是否身高有所不同？是否身高随年龄的增长呈线性关系？如果解决这样的问题，只建立一个控制层就不够了。应该考虑，

选择身高  $h$  作为因变量，分类变量  $age$  作为第一层控制变量， $sex$  为第二层控制变量。两个分类变量分别放在两层中，且使用选项。操作如下：

① 按前面叙述的方法先将变量  $age$ ，送入 Independent List 框中建立一个控制层。单击 Next 按钮，在变量表中选择第二个分类变量  $sex$ ，送入 Independent List 中，作为第二层。 $age$  和  $sex$  分别控制第一层和第二层。

② 单击 Options 按钮，展开 Options 对话框，见图 8-2。在 Statistics for First Layer 栏中选 Anova table and eta 和 Test of Linearity 两项。单击 Continue 按钮返回主对话框。

③ 运行的程序如下（在主对话框中单击 Paste 按钮，在 Syntax 窗口生成并运行）：  
MEANS

```
TABLES=h BY age BY sex
/CELLS MEAN STDDEV
/STATISTICS ANOVA LINEARITY .
```

与上一个程序比较可以看出在 MEANS 语句中，分类变量  $age$  作为第一 BY 变量，即第一层控制变量， $sex$  作为第二 BY 变量，CELLS 子命令指定只求因变量各单元均值和标准差，各单元指的是分别在两层中定义的两个分组变量各水平组合，STATISTICS 子命令指定 ANOVA 和 LINEARITY 对第一层变量进行方差分析和线性度检验。

④ 输出结果见表 8-3 至表 8-5。  
观察输出结果。与上一种分析对比，可以看出层变量的作用，还可以看出使用系统默认统计量和使用 Options 对话框中确定输出的统计量之间有什么不同。

表 8-3（已经转换盘处理）是由第一层变量  $age$  和第二层变量  $sex$  确定的各单元中身高均值。

表 8-3 各单元的身高均值表

Report						
身高	性别					
	女			男		
年龄	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
10	1.4500	.02000	1.4467	.02687	1.4488	.02167
11	1.5383	.02317	1.5000	.04637	1.5209	.03910
12	1.6100	.01414	1.6140	.01949	1.6129	.01704
13			1.5900		1.5900	
Total	1.5154	.06253	1.5357	.07623	1.5259	.06941

表 8-4 是方差分析与线性度检验的结果，说明如下：  
方差分析的变量信息：身高\*年龄，因变量  $h$  标签是“身高”，分组 BY 变量  $age$  标签为“年龄”。说明方差分析的要求是分析不同年龄的身高均值间是否存在显著性差异。  
表 8-4 中各统计量的名称与各统计量之间的数学关系：

- ① Sum of Squares 偏差平方和。
- ② Between Groups 组间偏差平方和。从表中可以看出组间偏差平方和 0.105（显示值，非机内值）。它由两部分组成：Linearity=0.097 是由因变量与控制变量之间的线性关系引起的，Deviation from Linearity=0.008 不是由因变量与控制变量之间的线性关系

引起的。

③ Within Groups=0.020 组内偏差平方和。各组内各观测相对于组均值的变异，有时也称其为误差变异。

④ Total 偏差平方和的总和。它等于组间偏差平方和 0.105 与组内偏差平方和 0.020 之和 0.125。df 是自由度，组间自由度 3，组内自由度 23。

表 8-4 对第一层变量的方差分析结果

ANOVA Table							
			Sum of Squares	df	Mean Square	F	Sig.
身高 * 年龄	Between Groups	(Combined)	.105	3	.035	39.587	.000
		Linearity	.097	1	.097	109.435	.000
		Deviation from Linearity	.008	2	.004	4.664	.020
	Within Groups		.020	23	.001		
	Total		.125	26			

表 8-5 关联度测度

Measures of Association				
	R	R Squared	Eta	Eta Squared
身高 * 年龄	.879	.772	.915	.838

⑤ Mean Square 均方。数值上等于偏差平方和除以自由度之商。

⑥  $F$  值，数值上等于组间均方与组内均方之比值。从表中可以看出组间偏差平方和 0.105、自由度 3、均方 Mean Square 值为组间偏差平方和除以自由度，值为  $0.105/3=0.035$ 。组内偏差平方和为 0.020、自由度 23、均方值为  $0.020/23=0.001$ （注意，因显示位数有限，此处是近似值）。

⑦ Sig，在四个年龄组学生身高均值相等的零假设下，获得各统计量的值或更极端值的概率。从表中可以看到，显著性概率近似为 0。组间均方远远大于组内均方，说明组间差异远远大于随机误差引起的组内差异。因此可得出结论：10、11、12、13 岁学生的身高差异显著。

线性回归方程的偏差平方和 0.097，均方 0.097， $F$  值为 109.437。Sig 近似为 0.000，即小于 0.001 说明回归方程预测性能很好。这也可以从  $R$  值为 0.879 接近 1 来说明。

表 8-5 关联度的测度：

⑧ Eta ( $\eta$ ) 值 0.915 说明因变量与自变量之间联系紧密。Eta squared ( $\eta^2$ ) 等于组间偏差平方和与总偏差平方和之比即  $0.105/0.125=0.84$ （表中因各项计算误差导致 0.838）。Eta 是 0~1 之间的数，越接近 1，就越表明因变量（身高）与控制变量（年龄）关系密切。如果 Eta=0 表明两个变量无关。

⑨  $R$  是因变量  $h$  观测值与预测值之间的线性相关系数，虽然没有直接求出回归方程，但可以知道， $R$  值越接近 1 表明线性回归方程的预测性能越好，因变量与自变量之间的线性回归关系越好。

⑩  $R^2$  是线性模型的拟合良好度，有时称作确定系数，是在因变量中由回归模型解



释的方差比例，其值的范围是 0~1。该值越小，表明模型对数据的拟合越不好。

#### 8.2.4 MEANS过程语句

使用下列命令语句和子命令调用 MEANS 过程如下：

```
MEANS [TABLES={varlist} BY varlist [BY...] [/varlist...]{ALL }  
[MISSING={TABLE**}{INCLUDE}{DEPENDENT}]  
[CELLS=[MEAN** ][COUNT** ][STDDEV**][MEDIAN]  
[GMEDIAN] [SEMEAN] [SUM ][MIN][MAX] [RANGE]  
[VARIANCE][KURT] [SEKURT] [SKEW] [SESKEW]  
[FIRST][LAST][NPCT][SPCT][NPCT(var)][SPCT(var)]  
[HARMONIC][GEOMETRIC]  
[DEFAULT][ALL] [NONE] ]  
[/STATISTICS=[ANOVA][LINEARITY][ALL][NONE]]  
[/STATISTICS=[ANOVA] [{LINEARITY}] [NONE**]]{ALL}
```

标有\*\*的子命令一旦省略，系统自动按选择默认项进行分析计算。默认项标有双星或使用黑体字。

MEANS 语句是调用 MEANS 过程的语句，“/”后面的是子命令，“[ ]”中的子命令是可以选择的。子命令中的选项分为两类：

使用 “[ ]”括起来的是复选项。即可以同时选择若干个在 [ ] 中的选项。

使用 “{ }”括起来的选项是单选题，只能择其一。

##### 1. MEANS 语句

MEANS 语句是调用 MEANS 过程的语句，可以在该语句中指定因变量和作为自变量的分类变量，也可以使用子命令形式为 MEANS 语句指定因变量和分类变量。

(1) TABLES=varlist 是指定因变量的方式，可以使用 “TABLES=”，也可以省略它，直接在 MEANS 关键字后面列出因变量的变量表，至少指定一个因变量。

BY 后面必须是分类变量。MEANS 过程按 BY 变量的值分组对因变量进行分析。

可以有不止一个 BY 分语句，但必须有至少一个 BY 语句。每个 BY 语句定义一个控制层，一个控制层中可以指定若干个分类变量作为层控制变量，有几个 BY 分语句就有几个控制层。控制层数（BY 分语句数目）和每层中的分类变量数目，以及每个分类变量的水平数决定观测量如何分组。

(2) 为 MEANS 语句提供因变量表和 BY 变量的另一种方法是使用子命令方式，即在 “/” 后面直接列出因变量表，并且紧接着各 “BY” 后面列出各层的分类变量。

##### 2. MISSING 子命令

MEANS 过程对自变量的缺失值作为该自变量的一个水平，单分一组给出统计量。

使用 MISSING 子命令指定处理因变量缺失值的方法，共有 3 个选项：

(1) Table 是系统默认的对缺失值的处理方法。对在 Table=后指定的任意一个变量带有缺失值的观测都会被从分析中剔除。这样, 包括在一个表中的每个观测都有一个对所有变量来说非缺失值的全集。当使用 “/” 分割一个表时, 每个表对缺失值分别处理。

(2) Include 将读者定义的缺失值当作合法值处理, 参与分析。

(3) Dependent 仅剔除因变量的读者缺失值。所有控制变量缺失值都作为合法值处理。

### 3. CELLS 子命令

CELLS 子命令指定对由 BY 变量确定的分析单元计算哪些统计量。可供选择统计量关键字如下:

(1) DEFAULT 所有系统默认的统计量, 包括均值、标准差、单元内观测量数目。

(2) MEAN、STD DEV、COUNT 分别为单元的均值、标准差、单元内的观测量数, 都是系统默认的统计量。不使用 DEFAULT, 可以使用这些关键字分别指定。

(3) SUM、VARIANCE、MEDIAN、GMEDIAN、SEMEAN 分别为单元内的因变量值的总和、单元的方差、单元的中位数、各组中位数、单元均值的标准误。

(4) MIN、MAX 和 RANGE 为单元中的最小值、最大值和范围。

(5) KURT、SEKURT 单元峰度和峰度的标准误。

(6) SKEW 和 SESKEW 单元偏度和偏度的标准误。

(7) FIRST、LAST 各分组第一个和最后一个观测的因变量的值。如果因变量指定的是个字符型变量, 输出只能给出这两个统计量和  $N$ 。

(8) NPCT、SPCT 分别为每组观测量数占总数的百分比、每组因变量总和占总和的百分比。

(9) [NPCT(var)]、[SPCT(var)]对指定变量求每组观测量数占总数的百分比、每组因变量总和占总和的百分比。

(10) HARMONIC、GEOMETRIC 分别为调和平均数、几何平均数。

(11) ALL、NONE 分别指定计算以上所有描述统计量、不计算描述统计量。

### 4. STATISTICS 子命令[/STATISTICS=[ANOVA][LINEARITY][ALL][NONE] ]

该子命令指定对第一层分类变量进行的统计分析, 可以指定以下选项:

(1) ANOVA 对第一层变量进行单变量方差分析。

(2) LINEARITY 当第一层变量的水平数大于等于 3 时, 对第一层变量进行线性度测量。指定此项, 会给出因变量观测值与预测值之间的相关系数和对线性回归的方差分析的假设检验结果。

(3) ALL、NONE 分别指定选择以上两项分析、对第一层变量不作特殊分析。

程序举例:

```
MEANS TABLES=V1 TO V5 BY GROUP
```

```
/STATISTICS=ANOVA.
```

该程序要求以  $V1$ 、 $V2$ 、 $V3$ 、 $V4$ 、 $V5$  为因变量, GROUP 为自变量分组进行均值比

较。省略了 CELL 子命令，相当于要求计算均值、标准差、各分组中观测量数。程序要求进行一维方差分析。

## 8.3 单一样本T检验

### 8.3.1 单一样本T检验的概念

One-Sample T Test 过程检验单个变量的均值是否与给定的常数之间存在差异。例如研究人员可能想知道一组学生的 IQ 平均分数与 100 分的差异。

如果已知总体均数，进行样本均数与总体均数之间的差异显著性检验属于单一样本的 T 检验。

变量的样本均值为  $\bar{x}$ ，已知总体均值(或给定常数)为  $\mu_0$ ，检验的零假设是  $H_0: \bar{x} = \mu_0$ 。计算公式为

$$t = \frac{\bar{x} - \mu_0}{s_x^-}$$

式中， $s_x^- = \frac{s}{\sqrt{n}}$  是均值标准误， $s$  是变量的标准差。

One-Sample T Test 过程对每个检验变量给出的统计量有：均值、标准差和均值的标准误。该过程计算每个数据值与总体均值之间差的平均值，进行该差值为 0 的 T 检验及计算该差值的置信区间，读者可以指定检验的显著性水平。

### 8.3.2 单一样本T检验的实例

**【例 2】**编号 data08-02 中的数据 1973 年某市测量的 120 名 12 岁男孩身高资料。已知该地区 12 岁男孩平均身高为 142.5cm，问该市男孩身高与该地区平均身高有否差异。

1. 建立无效假设  $H_0$ ：假设某市 12 岁男孩身高与该地区 12 岁男孩身高平均值相等。
2. 建立数据集仅有一个变量 Height：12 岁男孩身高。
3. 按 Analyze→Compare Mean→One Sample T Test 顺序展开 One Sample T Test 单一样本 T 检验对话框，如图 8-3 所示。
4. 在对话框中将唯一的变量 Height 从源变量栏移至 Test Variable(s)框内。在 Test Value 框中将该地区 12 岁男孩平均身高 142.5 输入到 Test 后的矩形框中，如图 8-3 所示。
5. 单击 Options 按钮，打开选项对话框，如图 8-4 所示。Confidence Interval 选择系统默认值 95%，缺失值选择系统默认的 Exclude cases analysis by analysis 项。单击 Continue 按钮，返回主对话框。
6. 如果在主对话框中单击了 Paste 按钮，则在语句窗口中产生如下命令程序。

T-TEST

```
/TESTVAL=142.5  
/MISSING=ANALYSIS  
/VARIABLES=height  
/CRITERIA=CI (.95).
```

7. 在对话框中单击 OK 按钮，输出结果见表 8-6 和表 8-7。

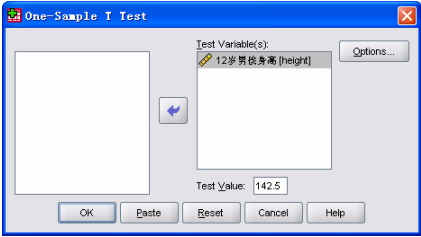


图 8-3 单一变量 T 检验对话框

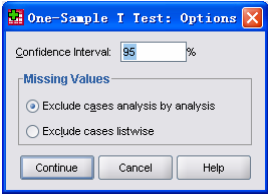


图 8-4 单一变量 T 检验选项

8. 结果分析

表 8-6 样本身高均值为 143.048，标准差 5.821，标准误 0.531。可以看出，样本均值 143.048 与地区身高平均值 142.5 比较，样本均值略高，差值为 0.548。

表 8-6 身高的基本描述统计量

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
12岁男孩身高	120	143.048	5.8206	.5313

表 8-7 单一样本 T 检验的分析结果

One-Sample Test						
Test Value = 142.5						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
12岁男孩身高	1.032	119	.304	.5483	-.504	1.600

表 8-7 中， $t$  值 1.032，自由度 119，双尾 T 检验的  $P$  值为  $0.304 > 0.05$ ，没有充分理由拒绝原假设。

Confidence Interval of the Difference 是差值的 95%置信区间。当总体标准差未知时，差值的 95%置信区间=均值差值 $\pm 1.96 \times$ 标准误。我们根据表 8-7 得知 95%置信区间是  $0.548 \pm 1.96 \times 0.531$ 。由此推出，95%置信区间为  $0.548 \pm 1.96 \times 0.531$ 。这就是表 8-7 中 Lower 与 Upper 两项中的数值 -0.504 和 1.600。这个 95%之心区间的含义是若以同样方式多次抽取等量的样本，对每个样本计算出的均值与总体均值的差异 95%落在这个区间之内。（注意：以上数值显示值与机内值有一定误差。因此如果使用计算器按显示值验算结果会稍有不同，大约误差小于 1%）。均值差值的 95%置信区间包括 0，没有充足理由拒绝样本均值与总体均值无显著差异的假设。

样本均值虽略高于总体均值，但无统计意义。误差来源可能是抽样误差，也可能来自测量误差。结论是，没有证据说明该市 12 岁男孩平均身高与该地区 12 岁男孩平均身高有显著性差异。

## 8.4 独立样本T检验

### 8.4.1 独立样本T检验的概念

进行独立样本的 T 检验要求被比较的两个样本彼此独立, 即没有配对关系。要求两个样本均来自正态总体。要求均值是对于检验有意义的描述统计量。

两个样本方差相等与不等时使用的计算  $t$  值的公式不同。因此应该先对方差进行齐性检验。SPSS 的输出, 在给出方差齐与不齐两种计算结果的  $t$  值, 以及 T 检验的显著性概率的同时, 还给出对方差齐性检验的  $F$  值和  $F$  检验的显著性概率。读者需要根据  $F$  检验的结果自己判断选择 T 检验输出中的哪个结果得出最后结论。

方差齐性检验的无效假设是: 两个独立样本来自方差相等的两个总体  $\nu_1 = \nu_2$ , 进行  $F$  检验。 $F$  值计算公式

$$F = \frac{\text{Max}(\nu_1, \nu_2)}{\text{Min}(\nu_1, \nu_2)}$$

式中,  $\nu_1$ 、 $\nu_2$  分别为两个样本的方差。两个方差较大的一个除以两个方差中较小的一个, 其比值为  $F$  检验的  $F$  值。

$p$  值小于 0.05 说明在该水平上否定原假设, 方差不齐。否则 ( $p$  值大于 0.05) 不足以在这个检验中拒绝原假设。(不排除在更多样本时, 或另一个检验方法时拒绝零假设)。

如果用  $\bar{x}_1$ 、 $\bar{x}_2$  表示两个样本的均值,  $n_1$ 、 $n_2$  分别为两个样本的观测值数目,  $\nu_1$ 、 $\nu_2$  为两个样本的方差, 方差齐 ( $\nu_1 = \nu_2$ ) 时与方差不齐 ( $\nu_1 \neq \nu_2$ ) 时计算  $t$  值使用的公式如下。

方差齐时公式

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

式中, 分母是两个样本均数之差的标准误; 其中  $S_c$  是合并方差, 公式为

$$S_c = \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

方差不齐时比较两个样本的均值, 可以对变量进行适当的变换使样本方差具有齐性, 再使用上述 T 检验计算公式进行计算与分析。SPSS 提供的函数可以实现对变量进行转换, 也可用下述公式计算  $t$  值并进行检验。在许多统计学书中称之为 T' 检验。SPSS 也在独立样本 T 检验过程的输出中提供方差不齐时使用下述公式计算的  $t$  值。

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\nu_1}{n_1} + \frac{\nu_2}{n_2}}}$$

独立样本 T 检验与配对样本 T 检验均使用 T Test 过程，但调用该过程的菜单不同，对数据文件结构的要求和所使用的命令语句也有区别。

### 8.4.2 独立样本T检验的过程

1. 按 Analyze→Compare Means→Independent Samples T Test 顺序，展开如图 8-5 所示的主对话框。
2. 源变量框中选择要进行检验的变量，将其送入 Test Variable(s)矩形框中。
3. 选择分组变量，将其送入 Grouping Variable)矩形框中，如图 8-5 所示。
4. 单击 Define Groups 按钮，展开 Define Groups 确定分组对话框，见图 8-6(a)和(b)。

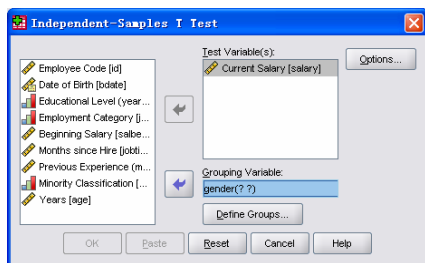
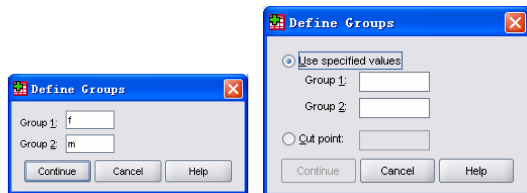


图 8-5 独立样本 T 检验主对话框



(a)

(b)

图 8-6 确定分类变量、连续变量的分组值

(1) 如果指定的 Grouping Variable 是分类变量（测度类型为 Nominal 或 Ordinal），且只有两个值，单击 Define Groups 按钮打开如图 8-6(a)的对话框，在 Group1 和 Group2 后面的矩形框中输入作为第一组和第二组的分类变量值。

(2) 如果指定的 Grouping Variable 是连续变量（测度类型为 Scale），或者测度类型为 Nominal 或 Ordinal 但有多值，Define Groups 按钮打开的对话框如图 8-6(b)所示。选择 Use specified values 使用指定的值，在 Group1 和 Group2 后面指定两个特定值，系统只对具有这两个值的因变量均值进行比较。

选择 Cut point 选项，在后面的矩形框中输入一个值。会将观测量按 Grouping Variable 值大于等于该值和小于该值分为两个组。检验在这两个组之间进行，比较其因变量在两组间的均值间是否差异显著。

#### 5. 选项对话框

在主对话框中，鼠标单击 Options 按钮，展开选项对话框，如图 8-4 所示。

(1) Confidence Interval 参数框。在该框中指定置信区间，系统默认值是 95%。可以在该项后面的文本框中重新输入一个由读者指定的百分比值。

(2) Missing Values 栏，选择对缺失值的处理方法。

① Exclude cases analysis by analysis，带有缺失值的观测量与分析有关时才被剔除。

② Exclude cases listwise 选项，剔除在 Test、Grouping 矩形框中的变量带有缺失值的观测量。

8.4.3 独立样本T检验的实例

【例 3】以银行男女雇员当前工资为例，见 data08-08，检验男女雇员当前工资是否有显著性差异。使用 gender 变量作为分类变量比较 salary 变量的均值。

首先假设银行雇员工资服从正态分布。检验的原假设 H0：不同性别果园的当前工资均值相等。取  $\alpha=0.05$ 。

(1) 读取数据文件 data08-08。按 Analyze→Compare Means→Independent Samples T Test 顺序，展开 Independent-Simple T Test 主对话框，如图 8-5 所示。按如下步骤操作，即可使用系统默认值进行检验。

(2) 选择 salary 作为检验变量，单击上面一个箭头按钮，将其送入 Test Variable(s)矩形框中。

(3) 选择 gender 变量作为分组变量，单击下面一个箭头按钮，将其送入 Grouping Variable 矩形框中，如图 8-5 所示。

(4) 单击 Define Groups 按钮，展开 Define Groups 确定分组的对话框。在 Group1 后面的矩形框中输入"f"为女雇员组；在 Group2 后面的矩形框中输入"m"，gender="m"，即男雇员作为第二组。

(5) 与选择系统默认值相应的命令语句

```
T-TEST
      GROUPS=gender ("f" "m")
      /MISSING=ANALYSIS
      /VARIABLES=salary
      /CRITERIA=CI (.95).
```

T-TEST 命令调用 T-test 过程，在该语句中只给出分组变量和分组方法。Groups=gender ("m" "f")指定分组变量为 gender，按其值 f、m 分为两组。

MISSING 子命令指定分析缺失值。VARIABLES=salary 指定对 salary 变量按 gender 分组进行分析。CRITERIA=CIN (.95)指定均值的置信区间为 0.95。

(6) 结果输出及结果说明，见表 8-8 和表 8-9。

表 8-8 分析变量的简单描述统计量

Group Statistics					
Gender		N	Mean	Std. Deviation	Std. Error Mean
Current Salary	Female	216	\$26,031.92	\$7,558.021	\$514.258
	Male	258	\$41,441.78	\$19,499.214	\$1,213.968

表 8-8 中是分析变量的简单描述统计量：

左第一栏为分析变量的标签 Current Salary（本例分析变量名 Salary）和分类变量标签 Gender 下方是用值标签表示的分组变量值，分为两组，一组为 female，另一组为 male。N 给出各组观测量数目，男 258 人，女 216 人。Mean 均值。给出各组观测量的分析变量均值。本例中男性雇员现平均工资为\$ 41441.78，女性雇员现平均工资为\$26031.92。分组给出分析变量的标准差 Std.Deviation，男组工资标准差为\$19499.214，女组工资标准差为\$7558.021。男组均值标准误 Std.Error Mean 为 1213.97，女组为 514.26。

表 8-9 给出方差齐性检验结果，以及 T 检验和校正 T 检验两种方法，并分别计算出的检验结果。

表 8-9 独立样本 T 检验的结果

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Current Salary	Equal variances assumed	119.669	.000	-10.945	472	.000	-\$15,409.862	\$1,407.906	-\$18,176.401	-\$12,643.322
	Equal variances not assumed			-11.688	344.262	.000	-\$15,409.862	\$1,318.400	-\$18,002.996	-\$12,816.728

① 方差齐性检验（Levene 检验）结果，F 值为 119.669，显著性概率为  $p<0.001$ ，因此结论是两组方差差异显著。在下面的 T 检验结果中应该选择 Equal variances not assumed，假设方差不相等一行的数据作为本例的 T 检验的结果数据。另一行 Equal variances assumed 是假设方差相等时 T 检验的计算结果。

② t 栏显示两个值。本例的 t 值等于-11.69。df 栏给出两种 T 检验的自由度。

③ Sig(2-Tailed)双尾 T 检验的显著性概率。本例的概率为 0.000，小于 0.05，否定不同性别雇员当前工资相等的原假设。可以得出结论男女雇员现工资具有显著差异。

④ Mean Difference 两组均值之差值为-\$15409.9。平均现工资女雇员低于男雇员 \$15409.9 元。

⑤ Std.Error Difference 差值的标准误为\$1318.40。

⑥ 95% Confidence interval of the Difference 差值的 95%置信区间。在\$-18003.0 ~ -12816.7 之间。不包括 0，也说明两组均值之差与 0 有显著差异。

结论：从 T 检验得 p 值为  $0.000<0.01$  和均值之差值的 95%置信区间不包括 0 都能得出，女雇员现工资明显低于男雇员，差异有统计意义。

注意，在实际应用中由于存在其他条件，如职务等级、工作经验等，不能得出平均工资差异是由性别差异造成的结论。根据分析结果得出结论要慎重。

【例 4】对连续变量按定点分组的独立样本 T 检验

现对 data08-03 数据进行独立样本 T 检验，有 29 名 13 岁男生的身高、体重、肺活量数据。试分析身高大于等于 155 厘米的与身高小于 155 厘米的两组男生的体重和肺活量均值是否有显著性差异。



首先建立无效假设  $H_0$ ：身高大于等于 155.0 与身高小于 155.0 两组之间的体重平均值在 99%水平上无显著差异，两组之间的肺活量平均值在 99%水平上无显著差异。

(1) 操作步骤

打开数据文件 data08-03。按 Analyze→Compare Means→Independent Samples T Test 顺序，展开 Independent-Simple T Test 独立样本 T 检验的主对话框。

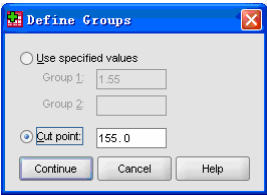


图 8-7 选项对话框

在源变量表中选择 weight、vcp 作为分析变量，用箭头按钮送入 Test Variable(s)框中。选择变量 height 作为分组变量，用箭头按钮送入 Grouping Variable 矩形框中。

单击 Define Groups 按钮展开定义分组对话框。选择 Cut point，并在后面的文本框中输入 155.0，见图 8-7。单击 Continue 按钮返回主对话框。

在主对话框中，单击 Options 按钮展开相应的对话框，见图 8-4。在 Confidence Interval 参数框中输入 99，单击 Continue 按钮返回主对话框。其他各选项使用系统默认值。

(2) 命令程序如下：

T-TEST

```
GROUPS=height(155.0)
/MISSING=ANALYSIS
/VARIABLES=weight vcp
/CRITERIA=CIN(.99).
```

(3) 运行结果与分析：结果输出见表 8-10 和表 8-11。

表 8-10 分组描述统计量

Group Statistics				
身高	N	Mean	Std. Deviation	Std. Error Mean
肺活量 >= 155.00	13	2.4038	.40232	.11158
< 155.00	16	2.0156	.42297	.10574
体重 >= 155.00	13	40.838	5.1169	1.4192
< 155.00	16	34.112	3.8163	.9541

表 8-11 方差齐性检验与 T 检验结果

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower Upper
肺活量	Equal variances assumed	.002	.961	2.512	27	.018	.38822	.15456	.07110 .70534
	Equal variances not assumed			2.525	2.6E1	.018	.38822	.15373	.07239 .70405
体重	Equal variances assumed	1.742	.198	4.056	27	.000	6.7260	1.6585	3.3231 10.1288
	Equal variances not assumed			3.933	2.1E1	.001	6.7260	1.7101	3.1771 10.2748

表 8-10 给出体重变量和肺活量变量按身高 height>=155.0 和 height<155.0 分组描述统计量。身高>=155.0 组 13 人，平均肺活量 2.4038，平均体重 40.838；身高<155.0 组 16

人平均肺活量 2.0156, 平均体重 34.113。

表 8-11 给出方差齐次性检验和 T 检验的计算结果。从 Sig 栏数据可以看出无论两组体重还是两组肺活量, 方差均是齐的, 均选择 Equal variances assumed 行的结果。

体重 T 检验的结果: Sig(2-tailed)=0.000 小于 0.001, 当然小于 1%拒绝原假设。从两组均值之差的 99%上下限均为正值, 也说明两组体重均值之差与 0 的差异显著。由此可以得出结论, 按身高 155.0 分组的两组体重均值差异, 在统计意义上高度显著。

肺活量 T 检验的结果: Sig(2-tailed)=0.018 大于 0.01, 从两组均值之差上下限一个为正值, 一个为负值也说明差值的 99%上下限与 0 的差异不显著。由此可以得出结论按身高 155.0 分组的两组肺活量均值差异在 99%水平上不显著。两组肺活量差值统计上不显著。均值差异是由抽样误差引起的。

## 8.5 配对样本T检验

### 8.5.1 配对样本T检验的概念

进行配对样本的 T 检验要求被比较的两个样本有配对关系, 要求两个样本均来自正态总体, 要求均值是对于检验有意义的描述统计量。均值的配对比较是比较常见的, 例如:

1. 同一窝实验用白鼠按性别、体重相同的配对, 再随机分到实验组和对照组, 分别喂加入海藻的饲料和普通饲料, 三个月后, 分别将每对白鼠置于水中, 测量其到溺死前的游泳时间。比较两组白鼠游泳时间均值, 从而比较两种饲料对抗疲劳的作用。

2. 同一组高血压病人, 在进行体育疗法前后, 测量其血压。每个病人在体育疗法前后的血压测量值构成观测量对。可以求这组病人体育疗法前后血压平均值。进行配对 T 检验, 分析体育疗法对降血压的疗效。

3. 在研究人体各部位体温是否有差别, 一个人的两个部位的温度构成一对数据。测量若干人的同样两个部位的温度数据, 可以比较这两个部位平均温度是否有显著性差异。使用配对 T 检验。

配对样本 T 检验实际上是先求出每对测量值之差值, 对差值变量求均值。检验配对变量均值之间差异是否显著。其实质检验的假设, 是差值变量的均值与零均值之间差异的显著性。如果差值均值与 0 均值无显著性差异说明配对变量均值之间无显著性差异。

如果差值变量为  $x$ , 差值变量的均值为  $\bar{x}$ , 样本的观测量数为  $n$ , 差值变量的标准差为  $S$ , 差值变量的均值标准误为  $S_{\bar{x}}$ , 配对样本 T 检验的  $t$  值计算公式为

$$t = \frac{\bar{x} - 0}{S_{\bar{x}}}, \quad S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

配对样本 T 检验与独立样本 T 检验均使用 T TEST 过程, 但调用该过程的菜单不同,

对数据文件结构的要求不同和所使用的命令语句也有区别。进行配对样本 T 检验的数据文件中一对数据必须作为同一个观测量中两个变量值。

## 8.5.2 配对样本 T 检验的过程

### 1. 建立数据文件

2. 按 Analyze→Compare Means→Paired-Samples T Test 顺序，展开 Paired-Samples T Test 配对样本 T 检验的主对话框，如图 8-8 所示。

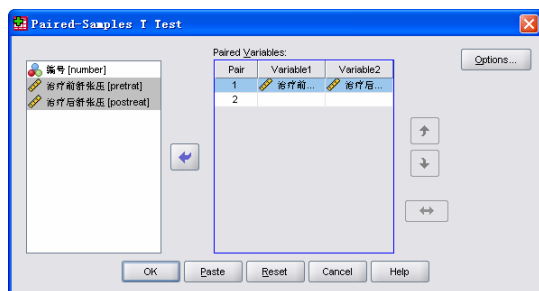


图 8-8 配对样本 T 检验主对话框

可以使用上述方法指定多个配对变量。以上操作是使用系统默认参数进行配对样本 T 检验的基本操作。单击 OK 按钮就可以提交运行了。

### 4. 配对样本 T 检验的选项

配对样本 T 检验使用系统默认值就可以得到比较满意的结果。如果想改变显著性概率，从而改变差值的置信区间，或者需要另外指定处理缺失值的方法，可以在主对话框中单击 Options 按钮，展开选项对话框，见图 8-4，在对话框中改变系统默认值，指定需要的选项。有关操作参见 8.4.2 小节，这里不再赘述。

## 8.5.3 配对样本 T 检验的实例

【例 5】现以体育疗法治疗高血压的数据为例，10 个高血压患者在施以体育疗法前后测定舒张压，数据编号 data08-04。数据文件中的变量：number 编号、pretreat 治疗前舒张压（mmHg）、postreat 治疗后舒张压（mmHg）。要求判断体育疗法对降低血压是否有效。这是一个自身配对样本的 T 检验问题。解决问题的步骤如下：

首先建立无效假设（ $H_0$ ）：体育疗法对高血压病人舒张压的降低无疗效，即对高血压病人治疗前后舒张压的差值均数是由差值为 0 的总体中随机抽取的，差值不为 0 是由抽样误差引起的。

1. 打开数据文件，按 Analyze→Compare Means→Paired-Samples T Test 顺序，展开 Paired-Samples T Test 配对样本 T 检验的主对话框，如图 8-8 所示。

### 3. 指定配对变量

① 在主对话框的源变量表中，选择一个变量，单击向右箭头按钮，变量名出现在 Paired Variables 框中 Variable1 列中；

② 在源变量框中，再选择一个与先选择的变量成对的变量，单击向右箭头按钮，变量名出现在 Paired Variables 框中 Variable2 列中。

## 2. 指定配对变量

配对变量为治疗前后的舒张压, 即 pretreat 和 postreat。在主对话框左面的变量表中, 单击 pretreat 变量, 按住 Ctrl 单击 postreat 变量, 鼠标单击向右的箭头按钮, 将配对变量送入 Paired Variables 矩形框中。单击 Continue 按钮, 确认并返回主对话框, 提交运行。

## 3. 运行的命令程序如下:

### T-TEST

PAIRS= pretreat WITH postreat (PAIRED)

/CRITERIA=CIN (.95)

/MISSING=ANALYSIS.

命令语句 T-TEST 调用 TTEST 过程。子语句 PAIRS...WITH...指定两个变量“(PAIRED)”说明 WITH 连接的两个变量是配对变量。CRITERIA=CIN 指定置信区间和显著性概率值。MISSING 子命令指定在分析时遇到带有缺失值的观测量时剔除该观测量。

## 4. 运行结果见表 8-12 至表 8-14。

表 8-12 治疗前后舒张压的简单描述统计量

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 治疗前舒张压	119.50	10	10.069	3.184
治疗后舒张压	102.50	10	11.118	3.516

表 8-13 治疗前后舒张压相关系数

	N	Correlation	Sig.
Pair 1 治疗前舒张压 & 治疗后舒张压	10	.599	.067

表 8-14 对配对变量差值的 T 检验

Paired Samples Test									
		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	治疗前舒张压 - 治疗后舒张压	17.000	9.534	3.015	10.180	23.820	5.639	9	.000

## 5. 结果分析

表 8-12 是对治疗前后舒张压的单变量描述统计量, 表中显示的是配对变量的变量标签, 对数为 1 对。

Means 给出治疗前后的舒张压均值, 分别为 119.50 与 102.50。N 即观测量数目治疗前后均为 10。Std. Deviation 治疗前后的舒张压的标准差, 分别为 10.069 和 11.118。

Std. Error Means 治疗前后的舒张压的均值标准误, 分别为 3.184 和 3.516。

表 8-13 中给出治疗前后舒张压之间的相关系数, Correlation 为 0.599, 不相关的概率为 0.067。相对于治疗前后舒张压的相关系数为 0 的假设成立的概率为 6.7%, 大于 5%, 可以得出结论是治疗前后的舒张压没有明显的线性关系。

表 8-14 给出配对变量差值的 T 检验结果。Means 均值之间的差值为 17.00; Std.

Deviation 差值的标准差为 9.53; Std. Error Means 差值的标准误为 3.01; 95% Confidence interval of the Deference 差值的 95%置信区间上下限为 10.18 和 23.82。应注意两个值均为正值。

t-value  $t$  值为 5.64; df 自由度为 9; Sig.(2-tailed)双尾 T 检验的结果, 获得  $t$  值的概率为 0.000, 即小于 0.001。

可以得出结论: 由于  $p$  小于 0.01。因此可以认为体育疗法对降低舒张压有明显疗效, 拒绝原假设。

## 8.6 T检验过程语句

T 检验的三种类型的检验都调用 T TEST 过程, 只是调用 T TEST 过程时的主语句格式不同。子命令都一样, 子命令顺序任意。

### 1. 主语句和必要的子命令

(1) 单一样本 T 检验的主语句和必要的子命令如下:

```
T-TEST TESTVAL n
```

①

```
/VARIABLE=varlist
```

②

① T-TEST 主语句调用 T-TEST 过程, 在 TESTVAL 后面给出一个标准值  $n$ 。

② VARIABLE 子命令后面列出要分析的变量。

TTEST 过程对 VARIABLE 子命令中的变量求均值, 并与 TESTVAL 后的标准值进行比较。T 检验的结果给出在 5%水平上差异的显著性概率。

(2) 独立样本 T 检验的主语句和必要的子命令如下:

```
T-TEST GROUPS=varname ({1,2**} {value} {value,value})
```

```
/VARIABLES=varlist
```

① T-TEST 是命令关键字。“GROUPS=”后面指定分组变量名, 并必须在后面的圆括号中给出分组方法。共有三种表示方法的选项:

- {1,2\*\*} 这种表示方法是针对常用情况的。分组变量值最常用的是 1 和 2。这是系统默认的分组变量值和分组方法。

- (value) 这是对连续变量作分组变量或分组变量虽然是分类变量但其值的数目多于 2 个的情况, 用针对分界值的方法分组时, 把分界值写入圆括号中。

- (value,value) 当分组变量值不是常用的 1 和 2 时, 或者分组变量值多于 2 个时, 指定两个分组变量值, 确定比较均值的分组。例如, 如果选择变量 AGE 作为分组变量, 其值有 11、12、13, 我们要进行比较的是 11 岁和 13 岁两组, 那么该命令语句写成 T-TEST GROUPS=(11,13)。如果分类变量是字符型, 括号中的两个字符各自加双引号。例如 T-TEST GROUPS=("f", "m")

② VARIABLES 子命令指定分析变量。等号后面跟着变量表, 对所列出的变量按

T-TEST 命令中的 GROUPS=子句给出的方法进行分组,求均值并检验两组均值的差异显著性。

### (3) 配对样本 T 检验的主语句

T-TEST PAIRS=varlist [WITH varlist [(PAIRED)]] [/varlist ...]

该语句要求在关键字 T-TEST 后面加“PAIRS=”表示后面跟着的是变量对。至少指定一对变量作为变量对。只有数值型变量可以被指定为变量对。变量对的表达方式如下:

① 格式 1: 变量表 WITH 变量表 (PAIRED) 简单配对格式。用 WITH 连接两个变量表,最后在圆括号中加“PAIRED”表明其前面的变量是配对的关系。配对关系必须是 WITH 前面的第一个变量与其后的第一个变量是一对,WITH 前面的第二个变量与其后的第二个变量是一对,以此类推。因此 WITH 前后的变量数目应该相等。否则多余的变量被忽略,并给出错误信息。书写时,要特别注意 WITH 前后变量的数目与顺序。若在表后给出 Paired,而没有在 PAIRS=后面加关键字 WITH,也会给出错误信息。

② 格式 2: PAIRS=varlist (即 PAIRS=变量表)。要求变量表中的每个变量都与变量表中的其他变量配对进行分析,例如: Pairs=A B C 则应该分析的变量对是 A 与 B, A 与 C, B 与 C。

③ 格式 3: 变量表/变量表,或 变量表 WITH 变量表。组合配对格式,这两种格式是相同的。WITH (或“/”)后面的每一个变量均与其前面的每一个变量配对。该格式不要求 WITH 前后变量数目相等。WITH 前面有  $m$  个变量,后面有  $n$  个变量,可以产生  $m \times n$  个变量对,例如: T-TEST PAIRS=A B C / D E 语句将进行 D-A、D-B、D-C, E-A、E-B、E-C 六个变量对进行 T 检验。

## 2. 子命令

下列子命令对所有类型的 T 检验都适用。

[/MISSING={ANALYSIS\*\*} {LISTWISE} {INCLUDE}]

[/CRITERIA=CI({0.95\*\*}) {value}]

带有\*\*的参数是该子命令不在程序中出现时的系统默认参数。

### (1) MISSING 子命令

① ANALYSIS 逐个分析地删除带有缺失值的观测量。这是系统默认的选择。

② LISTWISE 剔除在 PAIRS 子命令中指定的变量中带有缺失值的观测量。

③ INCLUDE 分析时包括读者定义的缺失值。读者缺失值作为一个合法值对待。

(2) CRITERIA=CI 子命令指定均值差值的置信区间,即检验的显著性概率水平,共有两种方式:

① 0.95 是系统默认值。

② value 读者自己给出具体数值。

## 习 题 8

1. 均值比较的 T 检验分几种类型？

2. 要使用 T 检验进行均值比较的变量，应该具有怎样的分布特征？

3. 两个独立样本 T 检验需要什么条件？

4. 一个品牌的方便面面饼的标称重量是 80 克，但是不能大小相差很大，因此要求标准差小于 2 克。现从生产线包装前的传送带上随机抽取部分面饼，称重数据记录在数据文件 data08-05。问这批面饼重量是否符合要求。

5. 某康体中心的减肥班学员入班时的体重数据和减肥训练一个月后的体重数据记录在数据文件 data08-06 中，试分析一个月的训练是否有效。如果按性别分组分析结果又如何？如果按体重等级分组检查训练效果，结果会是怎样的？

6. 为评价两个培训中心的教学质量，对两个培训中心学员进行了一次标准化考试，考试成绩如表中数据所示，分析两个培训中心教学质量是否有所差异？得出统计分析结果，并推断结论。数据 data08-07。

# 第9章 方差分析

## 9.1 方差分析的概念与方差分析过程

### 9.1.1 方差分析的概念

在科学实验中常常要探讨不同实验条件或处理方法对实验结果的影响。通常是比较不同实验条件下样本均值间差异。方差分析是检验多个样本均数间差异是否具有统计意义的一种方法,例如医学界研究几种药物对某种疾病的疗效,体育科研中研究训练目标、方法和不同运动量等因素对提高某项运动的成绩的效果,农业研究土壤、肥料、日照时间等因素对某种农作物产量的影响,不同饲料对牲畜体重增长的效果等,都可以使用方差分析方法去解决。

#### 1. 方差分析原理

方差分析的基本原理是认为不同处理组的均值间的差别基本来源有两个:

(1) 随机误差,例如测量误差造成的差异或个体间的差异,称为组内差异,用变量在各组的均值与该组内变量值之偏差平方和的总和表示,记做  $SS_w$ ,组内自由度记做  $df_w$ 。

(2) 实验条件或不同的处理造成的差异,称为组间差异。用变量在各组的均值与总均值之偏差的总平方和表示,记做  $SS_b$ ,组间自由度记做  $df_b$ 。例如,  $k \times m$  个实验对象随机分到  $k$  组,分别进行  $k$  种处理,要研究  $k$  种处理间均值是否存在显著差异,即处理是否有作用。测得数据如下,是单因素  $k$  水平的完全随机设计数据,见表 9-1。

表 9-1 单因素  $k$  水平的完全随机设计

	$j=处理 1$	处理 2	处理 3	处理 4	...	处理 $k$
$i=1$	$x_{11}$	$x_{21}$	$x_{31}$	$x_{41}$	...	$x_{k1}$
2	$x_{12}$	$x_{22}$	$x_{32}$	$x_{42}$	...	$x_{k2}$
3	$x_{13}$	$x_{23}$	$x_{33}$	$x_{43}$	...	$x_{k3}$
4	$x_{14}$	$x_{24}$	$x_{34}$	$x_{44}$	...	$x_{k4}$
5	$x_{15}$	$x_{25}$	$x_{35}$	$x_{45}$	...	$x_{k5}$
...	...	...	...	...	...	...
$m$	$x_{1m}$	$x_{2m}$	$x_{3m}$	$x_{4m}$	...	$x_{km}$

$i=1 \sim m$ , 是实验序号;  $j=1 \sim k$ , 是处理序号;  $x_{ij}$  是对第  $i$  个实验对象第  $j$  种处理后因变量测试值。

此为平衡设计,即各处理组实验对象数相等,均为  $m$  个。数据的完全随机分析中,可以证明,总偏差平方和分解为组间偏差平方和和组内偏差平方和之和:  $SS_t = SS_b + SS_w$ 。

总均值计算公式为



$$\bar{\bar{x}} = \frac{\sum_{j=1}^k \sum_{i=1}^m x_{ij}}{k * m}$$

第  $j$  种处理组均值为

$$\bar{x}_j = \frac{\sum_{i=1}^m x_{ij}}{m}$$

总偏差平方和如下, 其中  $x_{ij}$  是第  $j$  种处理组对第  $i$  个实验对象的观察值。

$$SS_t = \sum_{j=1}^k \sum_{i=1}^m (x_{ij} - \bar{\bar{x}})^2$$

组间偏差平方和如下, 反映处理间差异, 自由度  $df_b = k - 1$ 。

$$SS_b = m \sum_{j=1}^k (\bar{x}_j - \bar{\bar{x}})^2$$

组内偏差平方和如下, 总误差偏差平方和, 自由度  $df_w = k(m - 1)$ 。

$$SS_w = \sum_{j=1}^k \sum_{i=1}^m (x_{ij} - \bar{x}_j)^2$$

为去除样本量的影响,  $SS_b$ 、 $SS_w$  除以各自的自由度得到其均方值, 即组间均方和组内均方

$$MS_b = \frac{SS_b}{df_b}, \quad MS_w = \frac{SS_w}{df_w}$$

两者比值符合  $F$  分布, 自由度为  $(k - 1)$  和  $k(m - 1)$ 。

$$F = \frac{MS_b}{MS_w}$$

一种情况是处理没有作用, 即各样本均来自同一总体,  $MS_b/MS_w = 1$ 。考虑抽样误差的存在, 则有  $MS_b/MS_w \approx 1$ 。

另一种情况是处理确实有作用, 组间均方是由于误差与不同处理共同导致的结果, 即各样本来自不同总体。那么, 组间均方会远远大于组内均方, 即  $MS_b \gg MS_w$ 。

$MS_b/MS_w$  比值构成  $F$  分布。用  $F$  值与其临界值比较, 推断各样本是否来自相同的总体。

## 2. 方差分析的假定条件和假设检验

### (1) 方差分析的假定条件为:

- ① 各处理条件下的样本是随机的。
- ② 各处理条件下的样本是相互独立的, 否则可能出现无法解释的输出结果。
- ③ 各处理条件下的样本分别来自正态分布总体  $N(\mu_i, \sigma_i^2)$ , 否则使用非参数分析。
- ④ 各处理条件下的样本方差相同, 即具有齐性:  $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$

### (2) 方差分析的假设检验

假设有  $k$  个样本, 如果原假设  $H_0$ : 样本均数都相同即  $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k = \mu$ ,  $k$  个样

本有共同的方差  $\sigma^2$ ，则  $k$  个样本来自具有共同方差  $\sigma^2$  和相同均数  $\mu$  的总体。

如果经过计算，组间均方远远大于组内均方，则  $F > F_{0.05(df_b, df_w)}$ ， $p < 0.05$ ，推翻原假设，说明样本来自不同的正态总体，说明处理造成均值的差异有统计意义。否则， $F < F_{0.05(df_b, df_w)}$ ， $p > 0.05$ ，承认原假设，样本来自相同总体，处理间无差异。

### 9.1.2 方差分析中的术语

方差分析中常用的术语如下。

#### 1. 因素与处理

因素是影响因变量变化的客观条件，处理是影响因变量变化的人为条件，也可以统称为因素，实际上就是变量。例如影响农作物产量的气温、降雨量、日照时间等为因素。研究不同肥料对不同种系农作物产量的影响时农作物的不同种系可称为因素，所施肥料可视为不同的处理。一般情况下“因素”与“处理”在方差分析中可做相同理解。在要求进行方差分析的数据文件中均作为分类变量出现，即它们只有有限个取值。即使是气温、降雨量等平常看做是连续变量的，在方差分析中如果作为影响产量的因素进行研究，就应该将其数值用分组定义水平的方法事先变为具有有限个取值的离散变量。

#### 2. 水平

因素的不同等级称做水平。例如，性别因素在一般情况下只研究两个水平：男、女。化学或生物实验中的“剂量”必须离散化为几个有限的水平数，如 1ml、2ml、4ml 三个水平。

#### 3. 单元

在方差分析中单元 Cell 是指各因素的水平之间的每个组合。例如，研究问题中的因素有性别 Sex，取值为 0、1；有年龄，分三个水平 1（10 岁）、2（11 岁）、3（12 岁）。两个变量的组合共可形成六个单元：[1,1]、[1,2]、[1,3]、[2,1]、[2,2]、[2,3]，代表两种性别与三种年龄的六种组合。在方差分析中，比较各单元条件下，因变量均值之间的差异。

#### 4. 因素的主效应和因素间的交互效应

单独效应是在其他因素固定在某一水平时，因变量在某一因素不同水平间的差异。

因素的主效应就是因变量在一个因素各水平间的平均差异。

当一个因素的单独效应随另一个因素的变化而变化时，称两个因素间存在交互效应。

这是在科学实验和生产实践中常常遇到的问题。举例说明，有 A、B 两种药物治疗缺铁性贫血，患者 12 例，分为 4 组。实验方案是：第一组用一般疗法，第二组在一般疗法基础上加用 A 药，第三组在一般疗法基础上加用 B 药，第四组在一般疗法基础上 A、B 两种药同时使用，一个月后观察红细胞增加数，分析两种药物的疗效，数据见表 9-2。

数据来源于《医用统计方法》（金丕焕，人民卫生出版社）。

这是一个双因素方差分析的问题，因素 A 与因素 B。每个因素均有用该药与不用该药两个水平。研究药物 A 和 B 是否对红细胞的增加有显著影响需对红细胞增加数的均值作以下比较：

表 9-2 实验数据

	第一组	第二组	第三组	第四组
	红细胞增加数（百万/m <sup>3</sup> ）			
	0.8	1.3	0.9	2.1
	0.9	1.2	1.1	2.2
	0.7	1.1	1.0	2.0
各组平均值	0.8	1.2	1.0	2.1

(1) 比较第二组的均值与第一组的均值是否有显著性差异。

(2) 比较第三组的均值与第一组的均值是否有显著性差异。

这两项研究的是 A、B 两因素的主效应。

(3) 除了比较第四组的均值与第一组的均值是否有显著性差异外，还要研究 A 药对 B 药的疗效是否有影响。若 A 药对 B 药疗效无影响，那么除采样误差外，第四组与第二组均值之差应该等于第三组均值减去第一组均值。但是实际上  $(2.1-1.2)=0.9$ ， $(1.0-0.8)=0.2$ ，相差 0.7，该差值几乎与第一组均值相同。可以分析这个差异有统计意义，0.7 的差值包括采样误差和 A、B 药的相互作用。这种因素之间的相互作用在统计学上称为交互效应，在医学中称为协同效应（一个因素的单独效应随另一个因素的效应的增大而增大）或拮抗效应（一个因素的单独效应随另一个因素的效应的增大而减小）。如果交互效应存在，说明两个因素不是相互独立的。

5. 均值比较

均值的相对比较是比较各因素对因变量的效应大小的相对比较。例如研究 A、B 效应之和是否等于它们的交互效应，或者研究 A、B 对红细胞增加数的效应是否相等。

均值的多重比较是研究因素单元对因变量的影响之间是否存在显著性差异，例如例题中研究 A、B 药物对红细胞增加数的疗效是否存在显著性差异。

6. 单元均值、边际均值

在多因素方差分析中，每种因素水平组合的因变量均值称为单元均值。一个因素水平的因变量均值称为边际均值 (Marginal Means)，这是根据他们在表格中的位置命名的。见 9.3.4 小节中 2×2 析因方差分析例题中的解释。

7. 协方差分析

在一般进行方差分析时，要求除研究的因素外应该保证其他条件的一致。做动物实验往往采用同一胎的动物分组给予不同的处理，研究不同处理对研究对象的影响就是这个道理。例如，研究身高与体重的关系时要求按性别分别进行分析，这样消除性别因素的影响。不同年龄的身高与体重的关系也是有区别的，被测对象往往是不同年龄的。要消除年龄的影响，应该采用协方差分析。再如，研究几种饲料对增加动物体重的作用，以便比较哪种饲料更好，每个动物的进食量的影响应该在分析时消除，也需要进行协方差分析。

## 8. 重复测量

组内变异的主要原因是实验对象之间的个体差异。由于个体差异存在,即使实验对象受到相同的处理,它们的因变量值也可能相对不同。重复测量设计的方差分析也是像协方差分析一样,是在研究中减小个体差异带来的误差方差的一种有效方法,而且由于对相同个体进行重复测量在一定程度上降低了人力、物力、财力的消耗。

如果重复测量是在一段时间内或一个温度间隔内进行的,还可以研究因变量对时间、温度等自变量的变化趋势。这种重复测量研究称为趋势研究。例如将同一批动物在不同温度下生活一定时间并进行体重、脂肪的测定,可以研究时间、温度对动物体重、脂肪量的变化趋势的影响。

### 9.1.3 方差分析过程

SPSS 提供的方差分析过程包括:

#### 1. One-Way 过程

One-Way 过程是单因素的简单方差分析过程。它在 **Analyze** 菜单中的 **Compare Means** 过程组中,见图 9-1,用 **One-Way ANOVA** 菜单项调用。可以进行单因素的方差分析,在方差相等或不相等的情况下进行均值多重比较和详细的对比。

#### 2. General Linear Model (简称 GLM) 过程

GLM 过程由 **Analyze** 菜单直接调用。这些过程可以完成简单的多因素方差分析和协方差分析,不但可以分析各因素的主效应,还可以分析各因素间的交互效应。该过程允许指定最高阶次的交互效应,建立包括所有效应的模型。如果想建立包括某些特定的交互效应的模型也可以通过 **Model** 对话框中的选项实现。均值多重比较、绘制轮廓图等功能对比较各因素各水平的单元格均值,直观地判断因素间的交互效应非常有用。

在 **General Linear Model** 菜单项的下一级菜单中有 4 项,见图 9-2。每个菜单项分别完成不同类型的方差分析任务。

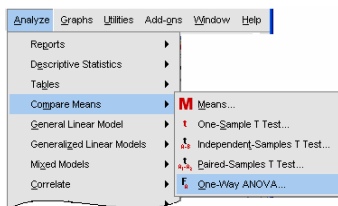


图 9-1 单因素方差分析的菜单图

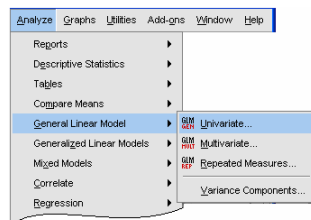


图 9-2 高级多元方差分析的菜单

#### (1) Univariate 过程

Univariate 过程提供回归分析和一个因变量与一个或几个因素变量的方差分析。因素变量把总体分为几组。使用这个一般线性模型过程,可以检验关于其他变量在单个因变

量各组均值效应的零假设。可以研究因素间交互效应以及单个因素（也可以是随机的因素）的效应。另外，还可以包括协变量效应和协变量与因素的交互效应。对回归分析，协变量指定作为自变量（预测变量）。在指定模型方面有较大的灵活性并可以提供大量的统计输出。

例如，如果以公司四个部门中的两个级别的职工为观察对象，研究生产率刺激机制，可以设计一个因子实验以便检验感兴趣的假设。由于在新刺激引入之前的原生产率可能对新刺激引入之后的生产率的比较发生很大影响，可以把原生产率作为协变量进行协方差分析。如果想看看协变量效应对两个级别的职工来说是否相同，也可以使用 **Univariate** 菜单项调用 **GLM** 过程进行分析。

### (2) Multivariate 命令

**Multivariate** 命令调用 **MANOVA** 过程进行多因变量的多因素分析。当研究的问题具有两个或两个以上相关的因变量，要研究一个或几个因素变量与因变量集之间的关系时，才可以选用 **Multivariate** 菜单项，例如研究数学、物理的考试成绩是否与教学方法、学生性别，以及方法与性别的交互作用有关时，使用此菜单项。如果只有几个不相关的因变量或只有一个因变量，应该使用 **Univariate** 菜单项调用 **GLM** 过程。

多因变量的多因素分析过程同样可以研究因素间交互效应以及单个因素的效应，该因素可以是随机的因素。另外，还可以包括协变量效应和协变量与因素的交互效应。对回归分析，自变量（预测变量）指定为协变量，可以检验平衡和不平衡模型。

### (3) Repeated Measures 命令

**Repeated Measures** 命令调用 **GLM** 过程进行重复测量方差分析。当一个因变量在同一课题中在不只一种条件下进行测量，要检验有关因变量均值的假设应该使用该过程。如果指定了被试间因素，他们把总体划分成几个组，可以检验组间因素的效应和组内因素的效应的零假设，可以检验单个因素的效应以及因素间的交互效应，另外，还包括协变量效应，以及被试间因素与协变量之间的交互效应。

### (4) Variance Components 命令

**Variance Components** 调用 **GLM** 过程进行方差成分分析。通过计算方差估计值，可以帮助我们分析如何减小方差。

## 9.2 单因素方差分析

单因素方差分析也称作一维方差分析。它检验由单一因素影响的一个（或几个相互独立的）因变量，由因素各水平分组的均值之间的差异，是否具有统计意义，并可以进行两两组间均值的比较，称作组间均值的多重比较，还可以对该因素的若干水平分组中哪些组均值间不具有显著性差异进行分析，即一致性子集检验。

**One-Way ANOVA** 过程要求因变量属于正态分布总体。如果因变量的分布明显的是

非正态，不能使用该过程，而应该使用非参分析过程。如果对被观测对象的实验不是随机分组的，而是进行的重复测量形成几个彼此不独立的变量，应该用 **Repeated Measures** 命令调用 **GLM** 过程对各因变量进行重复测量方差分析，条件满足时，还可以进行趋势分析。

### 9.2.1 简单的一维方差分析

【例 1】用四种饲料喂猪，共 19 头猪分为四组，每组用一种饲料。一段时间后称重，猪体重增加数据，见表 9-3。比较四种饲料对猪体重增加的作用有无不同。数据来源于《医用统计方法》（金丕焕，人民卫生出版社），数据编号 data09-01。

#### 1. 操作方法与步骤

(1) 在数据窗口中建立数据文件，定义两个变量，并输入数据，这两个变量如下：

**fodder** 变量，数值型，取值 1、2、3、4，分别代表 A、B、C、D 四种饲料。

**weight** 变量，数值型，其值为猪体重的增加数。

应该特别注意，不能把 A、B、C、D 定义为四个变量。

(2) 按 **Analyze→Compare Means→One-Way ANOVA** 顺序单击菜单，展开 **One-Way ANOVA** 主对话框，如图 9-3 所示。

表 9-3 饲料比较数据资料

饲 料			
A	B	C	D
133.8	151.2	193.4	225.8
125.3	149.0	185.3	224.6
143.1	162.7	182.8	220.4
128.9	143.8	188.5	212.3
135.7	153.5	198.6	

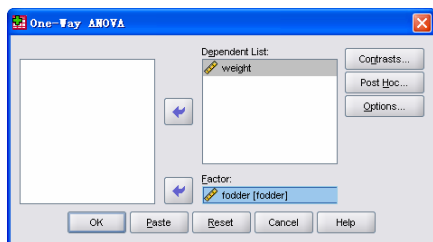


图 9-3 单因素方差分析的主对话框

(3) 根据分析要求指定方差分析的因变量和因素变量。

选定 **weight** 变量进入 **Dependent List** 框中，定义猪体重增加数为因变量。

选定 **fodder** 变量进入 **Factor** 框中，定义饲料为因素变量。

(4) 在主对话框中，单击 **OK** 按钮，输出窗口中可以见到如下的命令程序：

```
ONEWAY
```

```
Weight BY fodder
```

```
MISSING ANALYSIS.
```

**ONE-WAY** 主命令语句调用一维方差分析过程。**weight BY fodder** 是 **BY** 连接的两个变量，说明使用 **BY** 后面的变量水平将其前面的分析变量分组进行分析。**MISSING** 子命令说明对缺失值的处理方法。

#### 2. 输出结果见表 9-4

表 9-4 为因素变量饲料 **fodder** 对猪体重 **weight** 的影响分析结果。表的左上方是因变量 **weight**。

表 9-4 使用系统默认值的单因素方差分析结果

ANOVA					
weight					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20538.698	3	6846.233	157.467	.000
Within Groups	652.160	15	43.477		
Total	21190.858	18			

(1) 输出结果说明

第一栏：方差来源，包括组间偏差 Between Groups，组内偏差 Within Groups 和总偏差 Total。

第二栏：偏差平方和，组间偏差平方和为 20538.698，组内偏差平方和为 652.160，总偏差平方和为 21190.858，是组间偏差平方和与组内偏差平方和相加之和。

第三栏：自由度，组间自由度为 3，组内自由度为 15，总自由度为 18。

第四栏：均方，是第二栏与第三栏之比。组间均方为 6846.233，组内均方为 43.477。

第五栏： $F$  值，是组间均方与组内均方之比。

第六栏： $F$  值对应的概率值，针对假设  $H_0$ ：组间均值无显著性差异（即四种饲料对猪体重增加的平均值无显著性差异）。计算的  $F$  值 157.467，对应的概率值小于 0.001。

(2) 结果分析

根据输出的  $p$  值小于 0.001 可以看出，无论显著性水平取 0.05，还是取 0.01， $p$  值均小于临界值，因此否定  $H_0$  假设，四种饲料对猪体重的增加的均值差异显著。结论是四种饲料对猪体重的增加明显作用不同。

(3) 存在问题与解决方法

① 本例只考虑了猪体重的增加量，对其均值进行了比较，但实际工作中的问题往往不是这样简单，例如是否应该考虑每头猪的进食量对体重增加的影响，去除这个影响比较猪体重的增加会对饲料比较得出更切合生产实际的结论。这个问题应该使用 ANOVA 过程的协方差分析功能去解决。

② 使用系统默认值进行单因素方差分析只能得出是否有显著性差异的结论，本例数据量少，哪两组之间差别最大，哪种饲料使猪体重增加更快，几乎是可以看出来的。实际工作中往往需要两两的组间均值比较，这就需要使用 One-Way ANOVA 进行单因素方差分析时使用选项，从而获得更丰富的信息，使分析更深入。

③ 从主对话框可以看出 One-Way ANOVA 允许分析一个因素变量对多个因变量的影响，主对话框中的 Dependent List 栏中可以移入多个因变量。

9.2.2 单因素方差分析过程

单因素方差分析的选项分为三类：Contrasts 功能可以指定一种要用  $T$  检验来检验的 priori 对比；Post Hoc 功能可以指定一种多重比较检验；Options 功能可以指定要输出的统计量，指定处理缺失值的方法。分别使用主对话框中的三个按钮打开相应的对话框，

然后进行选择。

### 1. 进行对照比较的选项

在主对话框中,单击 **Contrasts** 按钮,打开对照比较对话框,如表 9-4 所示。在该对话框中可以把组间平方和划分成趋势成分或指定事先推测的对照比较。

#### (1) 趋势成分分析

考虑将组间偏差平方和分解为线性、二次、三次或更高次的趋势成分,操作如下:

① 选中 **Polynomial**,该操作激活其右面的 **Degree** 参数框。

② 单因素方差分析的 **One-Way ANOVA** 过程允许构造高达 5 次的均值多项式,多项式的阶数需要由读者自己根据研究的需要输入。单击 **Degree** 参数框右面的向下箭头展开阶次菜单,可以选择的阶次: **Linear** 线性、**Quadratic** 二次、**Cubic** 三次、**4th** 四次、**5th** 五次。系统将在输出中给出指定阶次和低于指定阶次的各阶的平方和分解结果和各阶次的自由度、*F* 值和 *F* 检验的概率值。



图 9-4 对照比较对话框

#### (2) 对照比较

① 系数指定规则。系数指定的顺序很重要,它应该与因素变量分组值的升序相对应。列表中第一个系数与因素变量最低组的值相对应,而最后一个系数与因素变量最高组的值相对应。例如,如果因素变量有六个水平,系数列为 1、0、0、0、0.5、0.5,对应着第一组到第六组。常用的是系数之和应该为 0,也可以设置系数之和不为 0,但会在输出中显示警告信息。表 9-4 中显示的是要求计算  $\text{mean1} \sim \text{mean4}$  的值,检验的假设  $H_0$ : 第一组均值与第四组的均值间无显著差异(差异无统计意义)。

② 指定各组均值的系数具体的操作步骤为:在 **Coefficients** 框中输入一个系数,单击 **Add** 按钮,**Coefficients** 框中的系数进入下面的方框中。重复上述操作,依次输入各组均值的系数,在方形显示框中形成一系列数值。因素变量有几个水平(分为几组),就输入几个系数,多出的无意义。不参与比较的分组系数应该为 0。如果多项式中只包括第一组与第四组的均值的系数,必须把第二个、第三个系数输入为 0 值。如果只包括第一组与第二组的均值,则只需要输入前两个系数,第三、四个系数可以不输入。

可以同时进行多组均值组合比较。一组系数输入结束,激活 **Next** 按钮,单击该按钮后 **Coefficients** 框被清空,准备接受下一组系数数据。最多可以输入 10 组系数。

如果认为输入的几组系数中有错误,可以分别单击 **Previous** 或 **Next** 按钮前后翻,找到出错的一组数据。单击出错的系数,该系数显示在编辑框中,可以在此进行修改,修改后单击 **Change** 按钮,在系数显示框中出现正确的系数值。当在系数显示框中选一个系数时,同时激活 **Remove** 按钮,单击该按钮将选中的系数清除。

### 2. 各组均值的多重比较选项

在主对话框中,单击 **Post Hoc** 按钮,展开 **Post Hoc** 多重比较对话框,如图 9-5 所示。



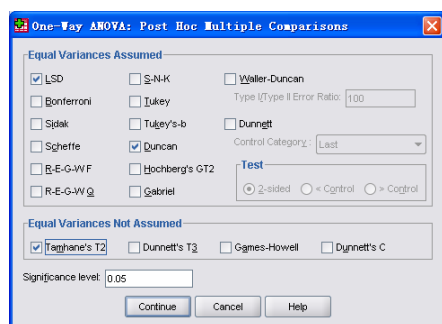


图 9-5 均值多重比较的对话框

在该对话框中提供近20种Post Hoc检验,有些检验均值差异是否显著,给出差异一致性子集,有些成对进行组均值比较,有些进行这两种检验。Post Hoc产生的多重比较检验表多达10种。非空组的组均值按升序排序,并用星号标明均值具有显著性差异的组对。另外,如果设计要求进行一致性子集检验,则计算一致性子集并将结果显示在一致性子集表中。

当各组观测量数目不同时,除R-E-G-WQ和R-E-G-WF外,在计算一致性子集时,使用各组

观测量数目的调和均值作为各组样本含量。而这两个成对比较的检验,都是用各组本身的样本含量。

### (1) 选择多重比较的方法

① 各组方差齐性时在 Equal Variance Assumed 栏中选择均值比较的方法,共 14 种方法,16 种选择。这些选项可以同时选择若干个,以便比较各种均值比较方法的结果。

- Least-significant difference (LSD), 用 T 检验完成各组均值间的配对比较,对多重比较误差率不进行调整。

- Bonferroni (LSDMOD), 计算 Student 统计量,完成各组间均值的配对比较。它通过设置每个检验的误差率来控制整个误差率。

- Sidak, 计算  $t$  统计量进行多重配对比较,调整多重比较的显著性水平。限制比 Bonferroni 检验更严格。

- Scheffe, 对所有可能的组合进行同步进入的配对比较。可以用于检验分组均数所有可能的线性组合。

- R-E-G-WF (Ryan-Einot-Gabriel-Welsch F), 用基于 F 检验的逐步缩小的多重比较检验,显示一致性子集表。

- R-E-G-WQ (Ryan-Einot-Gabriel-Welsch range test), 使用基于学生化值域逐步缩小的多元统计过程,进行子集一致性检验。

- S-N-K (Student Newman-Keuls), 使用学生化值域统计量,进行子集一致性检验。检验按均值递减排序,差异最大的先检验。

- Tukey (Tukey's honestly significant difference), 用 Student-Range 统计量进行所有组间均值进行配对比较,用所有配对比较的累计误差率作为实验误差率,还进行子集一致性检验。

- Tukey's-b, 用学生化极差统计量进行组间均值的配对比较,其精确值为前两种检验相应值的平均值。

• **Duncan (Duncan's multiple range test)**, 指定一系列的 Range 值, 逐步进行计算比较得出结论, 显示一致性子集检验结果。

• **Hochberg's GT2**, 是基于学生化最大模数的检验。与 Tukey 类似, 进行组均值成对比较和检测一致性子集。除非单元格含量非常不平衡, 该检验甚至适用于方差不齐的情况。

• **Gabriel**, 该方法根据学生化最大模数进行均值多重比较和子集一致性检验。当单元格含量不等时该方法比 Hochberg's GT2 更有效, 在单元格含量较大时, 这种方法较自由。

• **Waller-Duncan**, 用  $t$  统计量进行子集一致性检验。使用贝叶斯逼近。

• **Dunnett**, 使用 T 检验进行各组均值与对照组均值的比较。指定此选项, 进行各组与对照组的均值比较, 默认的对照组是最后一组。选择该项将激活下面的 **Control Category** 参数框。展开下拉列表, 可以重新选择对照组。在被激活的 **Test** 栏中选择是进行双尾 T 检验 (2 Sided)、各组均值是否都比对照组均值大的单尾 T 检验 (>Control), 还是各组均值是否都比对照组均值小的单尾 T 检验 (<Control)。

② 各组方差不具有齐性时, 在 **Equal Varance Not Assumed** 栏中选择检验各均数间是否有差异的方法, 有四种可供选择:

• **Tamhane's T2**, 用 T 检验进行各组均值配对比较。

• **Dunnett's T3**, 用学生化最大模数检验进行各组均值间的配对比较。

• **Games-Howell**, 进行各组均值配对比较检验, 该方法较灵活。

• **Dunnett's C**, 用学生化值域检验进行组均值配对比较。

③ 为便于选择, 下面按功能列出:

• 进行均值多重比较的选项有: **LSD**、**Sidak**、**Bonferroni**、**Games-Howell**、**Tamhane's T2**、**Dunnett's T3**、**Dunnett's C**、**Dunnett** (双尾、>Control、<Control)。

• 子集一致性检验的选项有 **SNK**、**Tukey's-b**、**Duncan**、**R-E-G-WQ**、**R-E-G-W F**、**Waller-Duncan**。

• 进行均值多重比较和子集一致性检验两种检验的选项有: **Hochberg's GT2**、**Tukey**、**Scheffe**、**Gabriel**。

(2) **Significance level** 选项设定各种检验的显著性概率临界值, 默认值为 0.05, 可由读者重新设定。

### 3. 输出统计量的选择

在主对话框中, 单击 **Options** 按钮, 展开选项对话框, 如表 9-6 所示。系统会按选择产生要求的统计量, 并按要求的方式显示这些统计量。在该对话框中还可以选择对缺失值的处理要求。各组选项的含义如下:

① **Statistics** 栏, 输出统计量的选项。

• **Descriptive**, 要求输出描述统计量。选择此项, 会计算并输出观测量数目、均值、标准差、标准误、最小值、最大值、各组中每个因变量的 95% 置信区间。

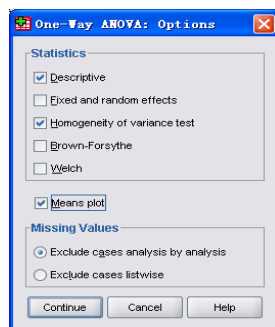


图 9-6 选项对话框

- **Fixed and random effects**, 输出固定效应模型的标准差、标准误和 95%置信区间, 以及随机效应模型的标准误、95%置信区间和方差成分间估测值。

- **Homogeneity of variance test**, 要求进行方差齐性检验, 并输出检验结果。用 Levene 检验计算每个观测值与其组均值之差, 然后对这些差值进行一维方差分析。

- **Brown-Forsythe**, 该统计量检验各组均值相等的, 当不能确定方差齐性假设时, 该统计量优于  $F$  统计量。

- **Welch**, 该统计量检验各组均值相等, 当不能确定方差齐性假设时, 该统计量优于  $F$  统计量。

② **Means plot** 栏, 要求做均值分布图, 根据因素变量值所确定的各组均值描绘出因变量的均值分布情况。

③ **Missing Values** 栏, 选择缺失值处理方法。

- **Exclude cases analysis by analysis**, 只有被选择参与分析的变量含缺失值的观测量从分析中剔除。

- **Exclude cases listwise**, 对所有含有缺失值的观测量从分析中剔除。

以上三组选项选择完成后, 按 **Continue** 按钮, 确认选择并返回主对话框。单击 **Cancel** 按钮取消本次选择返回主对话框。单击 **Help** 按钮, 显示有关的帮助信息。

### 9.2.3 单因素方差分析实例

【例 2】分析不同饲料对猪体重的影响数据 data09-01。

(1) 按 **Analyze**→**Compare Means**→**One-Way ANOVA** 顺序展开主对话框。

(2) 指定因变量: **weight** (体重); 因素变量: **fodder** (饲料)。

(3) 指定选项:

① 单击 **Contrasts** 按钮, 打开相应的对话框, 在 **Contrast** 栏中指定了 2 组系数:

1、0、0、-1。检验 A、D 饲料对猪体重增加的效应及其之间是否有显著性差异。

0.5、-0.5、0.5、-0.5。检验 A、C 饲料之和效应是否与 B、D 之和效应有显著差异。

② 单击 **Post Hoc** 按钮, 展开 **Post Hoc** 多重比较对话框, 见图 9-5, 选择均值多重比较的方法: 在 **Equal Variance Assumed** 栏中, 选择 **LSD**、**Duncan** 两种方法; 在 **Equal Variance Not Assumed** 栏中, 选择 **Tamhane's T2** 方法; 在 **Significance level** 框中, 选择 0.05。

③ 单击 **Options** 按钮, 展开选项对话框, 见图 9-6, 选择输出统计量: 选中 **Descriptive**, 要求输出描述统计量; 选中 **Homogeneity of variance test**, 做方差齐性检验; 选中 **Means plot**, 做均值分布图; 选中 **Exclude cases analysis by analysis**, 剔除参与分析的变量中有缺失值的观测量。

## (4) 命令语句

以上选择完成后在主对话框中单击 Paste 按钮, 在 Syntax 窗口中显示 ONEWAY 过程的命令语句如下:

```

ONEWAY                                     ①
weight BY fodder                           ②
/ CONTRAST= 1 0 0 -1 /CONTRAST=0.5 -0.5 0.5 -0.5  ③
/STATISTICS DESCRIPTIVES HOMOGENEITY          ④
/PLOT MEANS                                  ⑤
/MISSING ANALYSIS                           ⑥
/POSTHOC=DUNCAN LSD T2 ALPHA(.05).           ⑦

```

关于命令语句的说明

- ① 一维方差分析过程名, 调用一维方差分析过程进行单因素方差分析。
- ② 指定因变量为 **weight**, 按因素变量 **fodder** 的值分组。
- ③ 定义两组对比的系数。
- ④ 要求计算和输出描述统计量和进行方差齐性检验。
- ⑤ 绘制均数散点图。
- ⑥ 缺失值按分析要求进行个别剔除。
- ⑦ 各组均值的多重比较选择 **LSD**、**Duncan**、**Tamhane's T2** 方法。显著性概率临界值设定为  $\alpha=0.05$ 。

(5) 输出结果见表 9-5 至表 9-11、图 9-7。

表 9-5 描述统计量

Descriptives								
weight								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
A	5	133.3600	6.80794	3.04460	124.9068	141.8132	125.30	143.10
B	5	152.0400	6.95723	3.11137	143.4015	160.6785	143.80	162.70
C	5	189.7200	6.35035	2.83996	181.8350	197.6050	182.80	198.60
D	4	220.7750	6.10594	3.05297	211.0591	230.4909	212.30	225.80
Total	19	171.5105	34.31137	7.87157	154.9730	188.0481	125.30	225.80

表 9-6 方差齐性检验结果

Test of Homogeneity of Variances			
weight			
Levene Statistic	df1	df2	Sig.
.024	3	15	.995

表 9-7 单因素方差分析结果表

ANOVA					
weight					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20538.698	3	6846.233	157.467	.000
Within Groups	652.160	15	43.477		
Total	21190.858	18			

表 9-8 对比系数

Contrast Coefficients				
Contrast	fodder			
	A	B	C	D
1	1	0	0	-1
2	.5	-.5	.5	-.5

## (6) 结果说明

表 9-5 为描述统计量结果, 给出了四种饲料分组的样本含量  $N$ 、因变量猪体重的平

均数 Mean、标准差 Std. Deviation、标准误 Std. Error、95%的置信区间、最小和最大值。

表 9-9 对比结果

Contrast Tests							
		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
weight	Assume equal variances	1	-87.4150	4.42321	-19.763	15	.000
		2	-24.8675	3.03956	-8.181	15	.000
	Does not assume equal variances	1	-87.4150	4.31164	-20.274	6.852	.000
		2	-24.8675	3.01398	-8.251	14.649	.000

表 9-10 LSD 法和 Tamhane' s T2 法进行均值多重比较的结果

Multiple Comparisons							
Dependent Variable: weight							
		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval		
	(I) fodder (J) fodder				Lower Bound	Upper Bound	
LSD	A	B	-18.68000*	4.17024	.000	-27.5687	-9.7913
		C	-56.36000*	4.17024	.000	-65.2487	-47.4713
		D	-87.41500*	4.42321	.000	-96.8428	-77.9872
	B	A	18.68000*	4.17024	.000	9.7913	27.5687
		C	-37.68000*	4.17024	.000	-46.5687	-28.7913
		D	-68.73500*	4.42321	.000	-78.1628	-59.3072
	C	A	56.36000*	4.17024	.000	47.4713	65.2487
		B	37.68000*	4.17024	.000	28.7913	46.5687
		D	-31.05500*	4.42321	.000	-40.4828	-21.6272
	D	A	87.41500*	4.42321	.000	77.9872	96.8428
		B	68.73500*	4.42321	.000	59.3072	78.1628
		C	31.05500*	4.42321	.000	21.6272	40.4828
Tamhane	A	B	-18.68000*	4.35318	.016	-33.7633	-3.5967
		C	-56.36000*	4.16353	.000	-70.8053	-41.9147
		D	-87.41500*	4.31164	.000	-103.1431	-71.6869
	B	A	18.68000*	4.35318	.016	3.5967	33.7633
		C	-37.68000*	4.21260	.000	-52.3109	-23.0491
		D	-68.73500*	4.35904	.000	-84.6022	-52.8678
	C	A	56.36000*	4.16353	.000	41.9147	70.8053
		B	37.68000*	4.21260	.000	23.0491	52.3109
		D	-31.05500*	4.16966	.001	-46.4051	-15.7049
	D	A	87.41500*	4.31164	.000	71.6869	103.1431
		B	68.73500*	4.35904	.000	52.8678	84.6022
		C	31.05500*	4.16966	.001	15.7049	46.4051

\*. The mean difference is significant at the .05 level.

表 9-6 为方差齐性检验结果。从 sig=0.995 得出  $p>0.05$ ，说明各组的方差在 $\alpha=0.05$ 水平上没有显著性差异，即方差具有齐性。

表 9-7 是方差分析结果。给出了组间、组内的偏差平方和、均方、 $F$  值和概率  $p$  值 (sig)。 $p<0.05$ ，各组间均值在 $\alpha=0.05$ 水平上有显著性差异。

表 9-8 为对比系数表，列出两组均值对比的系数。用以检查对比目的是否表达正确。

表 9-9 为均值对比结果，表中内容解释如下：

第一栏：按方差齐性和非齐性划分。前面表 9-5 已得出方差具有齐性的结论，所以选择方差齐性（Assume Equal Variance）一行的数据得出结论。

第二栏：结合表 9-8 和表 9-7 得出该栏数据。第一个对比检验的是 A 组和 D 组均值是否有显著性差异，两组均值之差为-87.415 为 A-D 的值；第二个对比值为-49.735，是  $0.5A-0.5B-0.5C+0.5D$  的计算结果，其中大写字母代表各组因变量均值。

第三栏：标准误。

第四栏：计算的  $t$  值，是第二栏与第三栏之比。

第五栏：自由度。

第六栏： $t$  值的概率。从概率值可以看出：Contrast 1,  $p < 0.05$ ; Contrast 2,  $p < 0.05$ 。因此饲料对猪体重增加的效应，A、D 效应均值之间在  $\alpha = 0.05$  水平上有显著性差异。而 A、C 之和效应与 B、D 之和效应之间有显著性差异。从 Value of Contrast 栏内值的符号和描述统计表中 Mean 栏内的数据不难得出各对比组均值之差。

表 9-10 是 LSD 法和 Tamhane's T2 法进行均值多重比较的结果。从选择比较方法处知 LSD 属于 Equal Variance Assumed 框的选项，从表 9-5 得知方差具有齐性，因此只需从 LSD 法结果作结论。比较结果说明，A 与 B、A 与 C、A 与 D、B 与 C、B 与 D、C 与 D 各组均值间均有显著性差异。表中“\*”标示的组均值在 0.05 水平上有显著性差异。

表 9-11 为一致性子集检验结果。第一栏列出 A、B、C、D 各组。第二栏列出各组观测量数。由于各组样本含量不等，计算均数用的是调和平均数的样本量，为 4.706。各组猪体重增加量的均值单独为一个子集，说明没有两组均值相等的情况。与多重比较结果一致。

图 9-7 是以因素变量 fodder 为横轴，以独立变量 weight 为纵轴绘制的均值散点图。可直观地看出各组均值的分布。

表 9-11 DUNCAN 法一致性子集检验结果

		weight			
		Subset for alpha = .05			
fodder	N	1	2	3	4
Duncan <sup>a,b</sup> A	5	133.3600			
B	5		152.0400		
C	5			189.7200	
D	4				220.7750
Sig.		1.000	1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.706.

b. The group sizes are unequal. The harmonic mean of the group sizes is

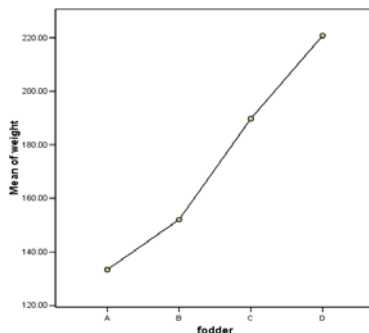


图 9-7 均值散点图

应该特别说明的是，选取哪些选项是根据研究需要进行的。本例中希望比较各种饲料对猪体重增加的效应，因此选择多重比较的选项。两个均值组合对比在此例中可能无实际意义，只是为了说明选项的使用方法才选择了 Contrast 选项。

【例 3】方差分析不同细菌对三叶草含氮量的影响。

本数据是 Erdman (1946) 的一个实验，同种三叶草被接种上不同的菌种测量三叶草植物中含氮量。每组数据中的前面一个是菌种代码，变量名是 strain，SPSS 分析过程要求因素变量必须为数值型变量。后面一个是含氮量，变量名是 nitrogen。数据编号为 data09-02。

下面是完成单因素方差分析的程序与输出结果。

(1) ONEWAY 命令语句（程序）

ONEWAY

nitrogen BY strain (1 30)

/STATISTICS DESCRIPTIVES HOMOGENEITY

/MISSING ANALYSIS

/POSTHOC=DUKEY LSD T2 ALPHA(0.05).

①

②

③

④

⑤

(2) 命令语句解释

从命令语句可知：①调用 ONEWAY 过程；②因变量 nitrogen，因素变量 strain，取值范围 1~30；③要求输出描述统计量和方差齐性检验的结果；④使用系统默认方法处理缺失值；⑤要求进行均值多重比较，采用 LSD、TUKEY、Tamhane's T2 方法；显著性概率临界值设定为 0.05。

(3) 输出结果见表 9-12 至表 9-16。

表 9-12 描述统计量结果分析表

Descriptives									
nitrogen									
	N	Mean	Std. Deviation	Std. Error	% Confidence Interval for Mean				
					Lower Bound	Upper Bound	Minimum	Maximum	
1	5	28.820	5.8002	2.5939	21.618	36.022	19.4	33.0	
4	5	14.640	4.1162	1.8408	9.529	19.751	9.1	19.4	
5	5	23.980	3.7772	1.6892	19.290	28.670	17.7	27.9	
7	5	19.920	1.1300	.5054	18.517	21.323	18.6	21.0	
13	5	13.260	1.4276	.6384	11.487	15.033	11.6	14.4	
30	5	18.700	1.6016	.7162	16.711	20.689	16.9	20.8	
Total	30	19.887	6.2422	1.1397	17.556	22.218	9.1	33.0	

表 9-13 方差齐性检验结果

Test of Homogeneity of Variances			
nitrogen			
Levene Statistic	df1	df2	Sig.
3.145	5	24	.025

表 9-14 单因素方差分析

ANOVA					
nitrogen					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	847.047	5	169.409	14.371	.000
Within Groups	282.928	24	11.789		
Total	1129.975	29			

此例输出与前面例题的输出格式一致，读者可以自己从中得出结论。需要注意的是表 9-13 方差齐性检验得出方差不具有齐性结论，在进行多重比较时应选择 Tamhane 方法作结论。从 Tamhane 方法的结果看，1 与 4，1 与 13，5 与 13，7 与 13，13 与 30 菌种之间的含氮量均值差异是有统计意义的。表 9-16 是指定 Tukey HSB 方法而产生的子集一致性检验结果，因为该方法要求方差具有齐性。

表 9-15 多重比较结果（一张表的三部分）

Multiple Comparisons									
Dependent Variable: nitrogen									
(I) strain (J) strain		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval				
					Lower Bound	Upper Bound			
Tukey HSD	1	4	14.1800*	2.1715	.000	7.466	20.894		
		5	4.8400	2.1715	.262	-1.874	11.554		
		7	8.9000*	2.1715	.005	2.186	15.814		
		13	15.5600*	2.1715	.000	8.846	22.274		
		30	10.1200*	2.1715	.001	3.406	16.834		
	4	1	-14.1800*	2.1715	.000	-20.894	-7.466		
		5	-9.3400*	2.1715	.003	-16.054	-2.626		
		7	-5.2800	2.1715	.185	-11.994	1.434		
		13	1.3800	2.1715	.987	-5.334	8.094		
		30	-4.0600	2.1715	.443	-10.774	2.654		
	5	1	-4.8400	2.1715	.262	-11.554	1.874		
		4	9.3400*	2.1715	.003	2.626	16.054		
		7	4.0600	2.1715	.443	-2.654	10.774		
		13	10.7200*	2.1715	.001	4.006	17.434		
		30	5.2800	2.1715	.185	-1.434	11.994		
	7	1	-8.9000*	2.1715	.005	-15.614	-2.186		
		4	5.2800	2.1715	.185	-1.434	11.994		
		5	-4.0600	2.1715	.443	-10.774	2.654		
		13	6.6600	2.1715	.053	-.054	13.374		
		30	1.2200	2.1715	.993	-5.494	7.934		
	13	1	-15.5600*	2.1715	.000	-22.274	-8.846		
		4	-1.3800	2.1715	.987	-8.094	5.334		
		5	-10.7200*	2.1715	.001	-17.434	-4.006		
		7	-6.6600	2.1715	.053	-13.374	.054		
		30	-5.4400	2.1715	.162	-12.154	1.274		
	30	1	-10.1200*	2.1715	.001	-16.834	-3.406		
		4	4.0600	2.1715	.443	-2.654	10.774		
		5	-5.2800	2.1715	.185	-11.994	1.434		
		7	-1.2200	2.1715	.993	-7.934	5.494		
		13	5.4400	2.1715	.162	-1.274	12.154		

LSD									
1	4	14.1800*	2.1715	.000	9.698	18.662			
	5	4.8400*	2.1715	.035	.358	9.322			
	7	8.9000*	2.1715	.000	4.418	13.382			
	13	15.5600*	2.1715	.000	11.078	20.042			
	30	10.1200*	2.1715	.000	5.638	14.602			
4	1	-14.1800*	2.1715	.000	-18.662	-9.698			
	5	-9.3400*	2.1715	.000	-13.822	-4.858			
	7	-5.2800*	2.1715	.023	-9.762	-.798			
	13	1.3800	2.1715	.531	-3.102	5.862			
	30	-4.0600	2.1715	.074	-8.542	.422			
5	1	-4.8400*	2.1715	.035	-9.322	-.358			
	4	9.3400*	2.1715	.000	4.858	13.822			
	7	4.0600	2.1715	.074	-.422	8.542			
	13	10.7200*	2.1715	.000	6.238	15.202			
	30	5.2800*	2.1715	.023	.798	9.762			
7	1	-8.9000*	2.1715	.000	-13.382	-4.418			
	4	5.2800*	2.1715	.023	.798	9.762			
	5	-4.0600	2.1715	.074	-8.542	.422			
	13	6.6600*	2.1715	.005	2.178	11.142			
	30	1.2200	2.1715	.579	-3.262	5.702			
13	1	-15.5600*	2.1715	.000	-20.042	-11.078			
	4	-1.3800	2.1715	.531	-5.862	3.102			
	5	-10.7200*	2.1715	.000	-15.202	-6.238			
	7	-6.6600*	2.1715	.005	-11.142	-2.178			
	30	-5.4400*	2.1715	.019	-9.922	-.958			
30	1	-10.1200*	2.1715	.000	-14.602	-5.638			
	4	4.0600	2.1715	.074	-.422	8.542			
	5	-5.2800*	2.1715	.023	-9.762	-.798			
	7	-1.2200	2.1715	.579	-5.702	3.262			
	13	5.4400*	2.1715	.019	.958	9.922			

Tamhane									
1	4	14.1800*	3.1807	.040	.569	27.791			
	5	4.8400	3.0954	.930	-8.690	18.370			
	7	8.9000	2.6427	.317	-6.522	24.322			
	13	15.5600*	2.6713	.044	.471	30.649			
	30	10.1200	2.6910	.206	-4.768	25.006			
4	1	-14.1800*	3.1807	.040	-27.791	-.569			
	5	-9.3400	2.4984	.083	-19.625	.945			
	7	-5.2800	1.9089	.485	-15.853	5.293			
	13	1.3800	1.9484	1.000	-8.860	11.620			
	30	-4.0600	1.9752	.769	-14.125	6.005			
5	1	-4.8400	3.0954	.930	-18.370	8.690			
	4	9.3400	2.4984	.083	-.945	19.625			
	7	4.0600	1.7632	.678	-5.535	13.655			
	13	10.7200*	1.8058	.026	1.444	19.996			
	30	5.2800	1.8348	.384	-3.839	14.399			
7	1	-8.9000	2.6427	.317	-24.322	6.522			
	4	5.2800	1.9089	.485	-5.293	15.853			
	5	-4.0600	1.7632	.678	-13.655	5.535			
	13	6.6600*	.8142	.001	3.250	10.070			
	30	1.2200	.8766	.968	-2.536	4.976			
13	1	-15.5600*	2.6713	.044	-30.649	-.471			
	4	-1.3800	1.9484	1.000	-11.620	8.860			
	5	-10.7200*	1.8058	.026	-19.996	-1.444			
	7	-6.6600*	.8142	.001	-10.070	-3.250			
	30	-5.4400*	.9595	.007	-9.398	-1.482			
30	1	-10.1200	2.6910	.206	-25.006	4.768			
	4	4.0600	1.9752	.769	-6.005	14.125			
	5	-5.2800	1.8348	.384	-14.399	3.839			
	7	-1.2200	.8766	.968	-4.976	2.536			
	13	5.4400*	.9595	.007	1.482	9.398			

\*. The mean difference is significant at the 0.05 level.

表 9-16 一致性子集检验结果

nitrogen					
		N	Subset for alpha = 0.05		
strain			1	2	3
Tukey HSD <sup>a</sup>	13	5	13.260		
	4	5	14.640		
	30	5	18.700	18.700	
	7	5	19.920	19.920	
	5	5		23.980	23.980
	1	5			28.820
	Sig.		.053	.185	.262

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 5.000.

## 9.2.4 单因素方差分析过程语句

### 1. ONEWAY 过程使用下列语句调用：

ONEWAY 因变量表 BY 因素变量（最小值，最大值）

[/POLYNOMIAL=n]

[/CONTRAST=系数表 1]

[/CONTRAST=系数表 2]

[/RANGE={LSD} {DUNCAN} {SNK} {TUKEY} {TUKEYB} {MODLSD} {SCHEFFE}

{.05} {alpha})



```
[/RANGE=...]  
[/STATISTICS={NONE} {DESCRIPTIVES} {EFFECTS} {HOMOGENEITY} {ALL}]  
[/POSTHOC={SNK}[TUKEY][BTUKEY][DUNCAN][SHEFFE][DUNNETT(refcat)]  
[DUNNETTL(refcat)][DUNNETTR(refcat)][BONFERRONI][LSD][SIDAK]  
[GT2] [GABRIEL] [FREGW] [QREGW] [T2] [T3] [GH][C]  
[WALLER ({100**} {kratio})] [ALPHA({0.05**} {Alpha})]  
[/MISSING={ANALYSIS**} {EXCLUDE} {INCLUDE} {LISTWISE}]  
[/MATRIX={IN (FILE)}[OUT (FILE)]
```

其中 ONEWAY 语句调用 ONEWAY 过程进行一维方差分析。因变量是等间隔变量, 因素变量是分类变量。在该主语句中必须指定因变量、因素变量, 并在因素变量后面的括号中给出因素变量的取值范围。其他各子命令均为选项。

因素变量作为自变量只能有一个。可以有最多达 100 个因变量, 自变量的水平数不限, 但均值多重比较 Post Hoc 检验要求自变量的非空水平数不能超过 50 个, Contrast 检验要求空与非空水平总数不能超过 50。

POLYNOMIAL 子命令只能有一个, POSTHOC 子命令也只能有一个, CONTRAST 子命令可以最多到 10 个。

## 2. 子命令

(1) POLYNOMIAL 子命令。组间偏差平方和分解为指定阶数的趋势成分。多项式阶数  $n$  取值范围为 1~5,  $n$  还必须小于自变量的分组数。该子命令只能在一个 ONEWAY 命令后面使用一次, 否则, 只有最后一个 POLYNOMIAL 子命令有效。

(2) CONTRAST 子命令。指定进行均值相对比较的子命令。该子命令指定均数比较时各水平的系数。CONTRAST 子命令后面的系数表中所列的系数顺序必须与数据文件中因素水平值顺序一致。一次调用 ONEWAY 命令可以同时指定多达 10 个 CONTRAST 子命令。每个 CONTRAST 子命令指定一个系数表。在调用语句中先出现的 CONTRAST 子命令指定的系数表, 在输出表中命名为 Construct1, 接下来的为 Construct2…。

(3) RANGE 子命令。在各组方差具有齐性的条件下使用 RANGE 子命令要求对均值进行一致性子集检验。后面 “{ }” 中的选项是进行检验的方法, 可以选择一种, 也可以同时选择几种。如果不指定显著性概率, 自动使用 0.05, 也可以由读者自己指定, 给出的概率数值要放在括号内。一个 RANGE 子命令指定一种检验的方法。在一次调用 ONEWAY 过程中进行几种一致性子集检验, 就必须使用几个 RANGE 子命令。指定一致性子集检验的方法必须按关键字原样照写, 不能使用缩写。

(4) STATISTICS 子命令。指定要求输出的统计量。可以选择的有:

① DESCRIPTIVES 要求输出描述统计量, 包括各组的均值、标准差、标准误、最大值、最小值、均值的 95% 上下置信区间。

② HOMOGENEITY 要求进行方差齐性检验。

③ **EFFECTS** 要求进行有效性检验。

④ **ALL** 要求输出以上各项统计量。

其中①、②两项是可以通过对对话框指定的,对话框中不包括的选项可以在对话框各项指定完成后,用 **Paste** 按钮,将程序语句生成在 **Syntax** 窗口中,再加以修改。

(5) **POSTHOC** 子命令。指定多重比较方法。等号后面是各种多重比较方法的关键字,关键字不能用缩写,必须原样照写。

(6) **MISSING** 子命令。指定缺失值的处理方法。

① **ANALYSIS** 参与计算的观测量中带有缺失值时,该观测量从有关的分析中剔除。此选项是系统默认的方法。

② **EXCLUDE** 把带有缺失值的观测量从分析中剔除。

③ **INCLUDE** 不剔除带有缺失值的观测量。

④ **LISTWISE** 所有在变量表中出现的变量,带有缺失值的观测量都从分析中剔除。

(7) **MATRIX** 子命令。指定输入的矩阵数据文件和输出的矩阵数据文件。指定输入矩阵文件名放在 **IN** 后面的括号中。指定输出矩阵文件,将文件名放在 **OUT** 后面的括号中。有关矩阵数据文件的生成程序请参见第 10 章“相关分析”一章的有关内容。

## 9.3 单因变量多因素方差分析

### 9.3.1 单因变量多因素方差分析概述

#### 1. 概述

单因变量多因素方差分析是对一个独立变量是否受多个因素或变量影响而进行的方差分析。**SPSS** 调用 **UNIANOVA** 过程,检验不同水平组合之间因变量均数由于受不同因素影响是否有差异的问题。在这个过程中可以分析每一个因素的作用,也可以分析因素之间的交互作用。可以进行协方差分析,以及各因素变量与协变量之间的交互作用。该过程要求因变量是从多元正态总体随机采样得来,且总体中各单元的方差相同,也可以通过方差齐性检验选择均值比较结果。

因变量和协变量必须是数值型变量,协变量与因变量彼此不独立。因素变量是分类变量,可以是数值型也可以是长度不超过 8 的字符型变量。固定因素变量 (**Fixed Factor**) 反应处理的因素。随机因素是随机设置的因素,是在确定模型时需要考虑会对实验有影响的因素,对实验结果影响的大小可以通过方差成分分析确定。

#### 2. 关于模型

**GLM Univariate** 功能很强,可以建立包括各种主效应、交互效应的模型。必须认真分析因素变量的具体情况,来确定自己的模型,否则会产生不可解释的输出结果。

9.3.2 单因变量多因素方差分析过程

单因变量多因素方差分析的功能模块调用步骤如表 9-2 所示。即按 Analyze→General Linear Model→Univariate 顺序单击菜单，展开 Univariate 主对话框，如图 9-8 所示。

用与 9.2 节中叙述的相同方法确定因变量，将因变量移到 Dependent Variable 框中。定义固定因素变量，并将其移到 Fixed Factor(s)框中。将随机因素变量移到 Random Factor(s)框中。

注意，由于内存容量的限制，选择的因素水平组合数（单元数）应该尽量少。因素数量和对选定因素定义的取值数量决定了组合数。

如果需要去除协变量的影响，将协变量移到 Covariates 框中。

WLS Weight 允许指定一个权重变量，用于加权的最小平方分析。权重变量给观测量不同的权重，也可以给不同测量精度以不同的补偿。如果需要考虑权重变量的影响，将权重变量移到 WLS Weight 框中。

可通过功能按钮展开相应对话框选择模型、对比和选择输出统计量。

1. 选择分析模型

在主对话框中，单击 Model 按钮，展开 Univariate: Model 对话框，见图 9-9。

(1) 在 Specify Model 栏中，指定模型类型。

① Full factorial 为系统默认的模型，即全模型。全模型包括所有因素变量的主效应、所有协变量主效应、所有因素与因素的交互效应，不包括协变量与其他因素的交互效应。不打开此对话框，即选择了全模型。

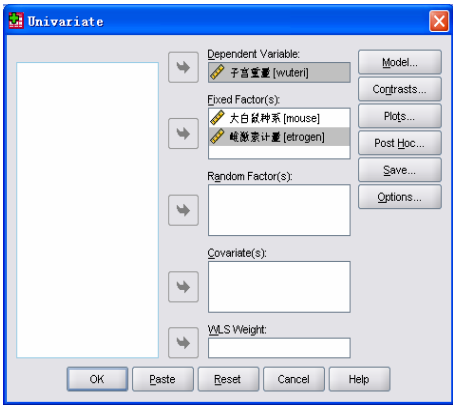


图 9-8 单因变量多因素方差分析主对话框

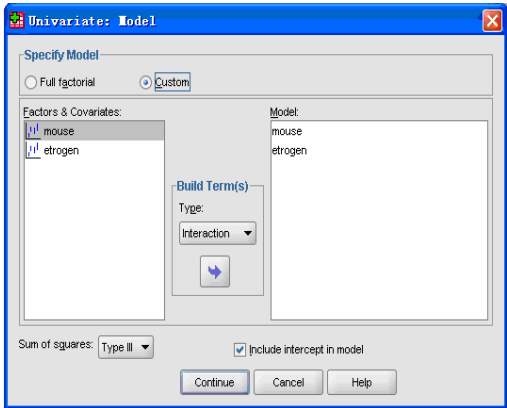


图 9-9 定义分析模型对话框

② Custom，建立自定义的模型。此项的选择激活下面各操作框。

(2) 建立自定义模型

选择了 Custom 后，在 Factors & Covariates 框中自动列出可以作为因素变量的变量名，根据表中列出的变量名建立模型。

### ① 选择模型中的主效应

选择一个因素变量名, 单击 **Build Term(s)** 栏中下面的箭头, 送入 **Model** 框中, 一个变量名占一行称为主效应项。欲在模型中包括几个主效应项, 就进行几次如上的操作。也可以选择多个一次送入模型框中。

### ② 选择交互效应类型

单击 **Build Term(s)** 右面的向下箭头可以展开小菜单。可以看到有如下几项:

- **Main effects**, 指定主效应。
- **Interaction**, 选中此项可以指定任意的交互效应。
- **All 2-way**、**All 3-way**、**All 4-way**、**All 5-way** 选项, 指定所有二维交互效应到所有五维交互效应。在下拉菜单中单击某一项, 选中的交互类型显示在矩形框中。

### ③ 建立模型中的交互项。以三个因素变量为例, 方法如下:

- 要求模型中包括两个变量的二维交互效应。相应的操作是在 **Factors & Covariates** 框内的变量表中选择一个变量, 此为选择了交互项之一, 再选择第二个变量, 此为选择了交互项之二。单击 **Build Term(s)** 栏内参数框的箭头按钮, 一个交互效应出现在 **Model** 框中。模型增加了一个交互效应项: 两个变量名之间用 “\*” 连接。

- 要求模型中包括三个变量的所有二维交互效应项时应该分别用鼠标单击三个变量名。在 **Build Term(s)** 栏内参数框中选择 **All 2-way** 项, 单击箭头按钮。在 **Model** 框中出现三个二维交互效应项: 两两变量名用星号连接的表达式共三个。

- 若要求模型中包括所有三维效应, 分三次单击三个变量, 选择 **Build Term(s)** 栏内参数框中的 **Interaction** 或 **All 3-way** 项, 再单击箭头按钮, 均可以在 **Model** 框中出现三维交互效应项: 三个变量名间用星号连接。

### (3) 选择分解平方和的方法

在对话框的下部有 **Sum of squares** 选项框, 可以进行四项选择来确定平方和的分解方法: **Type I**、**Type II**、**Type III** 和 **Type IV**。其中 **Type III** 是系统默认的, 也是常用的一种方法。

① **Type I**, 分层处理平方和的方法。仅对模型主效应之前的每项进行调整。一般适用于: 平衡的 ANOVA 模型, 在这个模型中一阶交互效应前指定主效应, 二阶交互效应前指定一阶交互效应, 依次类推; 多项式回归模型中, 任何低阶项都在较高阶项前面指定; 完全嵌套模型, 在模型中第一个被指定的效应嵌套在第二个被指定的效应中, 第二个被指定的效应嵌套在第三个被指定的效应中, 嵌套模型只能使用语句指定。

② **Type II**, 该方法计算一个效应的平方和时, 对其他所有的效应进行调整。一般适用于: 平衡的 ANOVA 模型、仅有主效应的模型、任何回归模型、完全嵌套设计。

③ **Type III**, 是系统默认的处理方法。对其他任何效应均进行调整。它的优势是把所估计剩余常量也考虑到单元频数中。一般适用于: **Type I**、**Type II** 所列的模型和没有空单元格的平衡和不平衡模型。

④ Type IV, 该方法是为有缺失单元格的情况设计的。使用此方法对任何效应  $F$  计算平方和。如果  $F$  不包含在其他效应里, Type IV=Type III=Type II。如果  $F$  包含在其他效应里, Type IV只对  $F$  的较高水平效应参数作对比。一般适用于: Type I、Type II 所列模型和有空单元格的平衡和不平衡模型。

(4) 选中 Include intercept in model, 系统默认截距包括在回归模型中。如果能假设数据通过原点, 可以不包括截距, 就不选择此项。

## 2. 选择对照方法

在主对话框中, 单击 Contrasts 按钮, 展开 Univariate: Contrasts 对话框, 如图 9-10 所示。

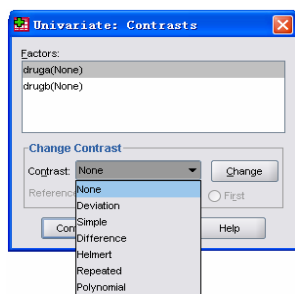


图 9-10 选择对照方法对话框

(1) 在 Factors 框中显示出所有在主对话框中选中的因素变量。因素变量名后的括号中是当前的对比方法。

(2) 在 Change Contrast 栏中改变对照方法。Contrast 检验一个因素的各水平间的差异。可以对模型中的每个因素指定一种对比方法, 对比结果描述的是参数的线性组合。操作方法如下。

① 在 Factors 框中选择想要改变对照方法的因子。激活 Change Contrast 栏中的各项。

② 单击 Contrast 参数框中的向下箭头, 在展开的对照方法表中选择对照方法, 可供选择的对照方法包括:

- None, 不进行均数比较。
- Deviation, 除被忽略的水平外, 比较预测变量 (或称因素变量) 的每个水平的效应。可以选择 Last (最后一个水平) 或 First (第一个水平) 作为忽略的水平。
- Simple, 除了作为参考的水平外, 对预测变量或因素变量的每一水平都与参考水平进行比较。选择 Last 或 First 作为参考水平。
- Difference, 对预测变量 (或因素) 的每一水平的效应, 除第一水平以外, 都与其前面各水平的平均效应进行比较。与 Helmert 对照方法相反。
- Helmert, 对因素的效应, 除最后一个以外, 都与后续的各水平的平均效应相比较。
- Repeated, 对相邻的水平进行比较。对因素的效应, 除第一水平以外, 对每一水平都与它前面的水平进行比较。
- Polynomial, 多项式比较。第一级自由度包括线性效应与预测变量或因素水平的交叉。第二级包括二次效应等。各水平彼此的间隔被假设是均匀的。

③ 单击 Change 按钮, 选中的 (或改变了的) 对照方法显示在步骤①选中的因子变量后面的括号中。

④ 只有选择了 Deviation 或 Simple 方法时才需要选择参考水平。共有两种参考水平可选择, 最后一个水平 Last, 和第一水平 First。系统默认的参考水平是 Last。

### 3. 选择分布图形

在主对话框中单击 **Plots** 按钮, 展开 **Univariate: Profile Plots** 对话框, 如图 9-11 所示。在该对话框中, 选择做边际均值图的参数。

边际均值图 (**Profile**) 用于比较边际均值。边际图是线图, 图中每个点表明因变量在因素变量每个水平上的边际均值的估计值。如果指定了协变量, 该均值则是经过协变量调整的均值。纵轴是因变量; 横轴是一个因素变量。

做单因素方差分析时, 边际图表明该因素各水平的因变量均值。

双因素方差分析时, 指定一个因素做横轴变量, 另一个因素变量的每个水平产生不同的线。

如果是三因素方差分析, 可以指定第三个因素变量, 该因素每个水平产生一个边际图。双因素或多因素边际图中的相互平行的线表明在因素间无交互效应, 不平行的线表明因素间存在交互效应。见图 9-12 和图 9-13。

(1) **Factors** 框中为主对话框中所选因素变量名。

(2) **Horizontal Axis** 框。选择 **Factors** 框中一个因素变量做横坐标变量, 单击箭头按钮, 将其送入相应的横坐标轴框中。

如果只想看该因素变量各水平的因变量均值分布, 单击 **Add** 按钮, 将所选因素变量移入下面的 **Plots** 框中, 否则, 不单击 **Add** 按钮, 接着做下一步。

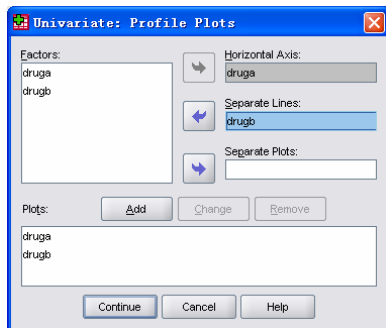


图 9-11 选择分布图形对话框

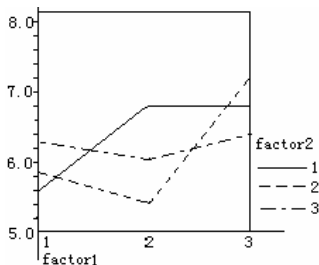


图 9-12 两因素变量有交互作用

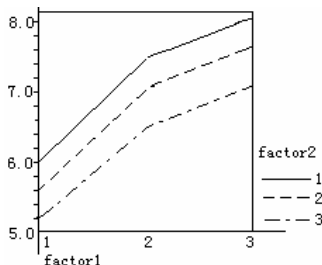


图 9-13 两因素变量无交互作用

(3) **Separate Lines** 框, 确定分线变量。如果想看两个因素变量组合的各单元格中因变量均值分布, 或想看两个因变量间是否存在交互效应, 选择 **Factors** 框中另一个因素变量, 单击箭头按钮, 将变量名送入 **Separate Lines** 框中。单击 **Add** 按钮, 将自动生成的图形表达式送入 **Plot** 栏中。分线框中变量的每个水平将在图中是一条线。图形表达式是用 “\*” 连接的两个因素变量名。

(4) **Separate Plots** 框, 确定分图变量。如果在 **Factor** 栏中还有因素变量, 可以按上述方法, 将其送入 **Separate Plot(s)** 框中, 单击 **Add** 按钮, 将自动生成的图形表达式送入 **Plot(s)**

栏中。图形表达式是用“\*”连接的三个因素变量名。分图变量的每个水平生成一张线图。

(5) 将图形表达式送到 Plots 框后发现有错误, 可以修改和删除。单击有错的图形表达式, 该表达式所包括的变量显示到输入的位置上。对选错的变量, 将其送回源变量框中。再重新输入正确内容。然后单击 Change 按钮改变表达式。在检查无误后, 按 Continue 按钮确认, 返回到主对话框。

#### 4. 选择多重比较分析

在主对话框中, 单击 Post Hoc 选项, 展开 Univariate: Post Hoc Multiple Comparisons for Observed 对话框。从 Factor(s) 框选择变量, 单击箭头键, 使被选变量进入 Post Hoc test for 框。然后选择多重比较方法。方法的选择请参见 9.2.2 节。

#### 5. 保存运算结果的选项

在主对话框中, 单击 Save 按钮, 展开 Univariate: Save 对话框, 如图 9-14 所示。通过在对话框中的选择, 系统使用默认变量名将所计算的预测值、残差值和诊断值作为新的变量保存在编辑数据文件中。以便于在其他统计分析中使用这些值。在数据编辑窗口中, 使用鼠标指向变量名, 会给出对该新生成变量含义的解释。

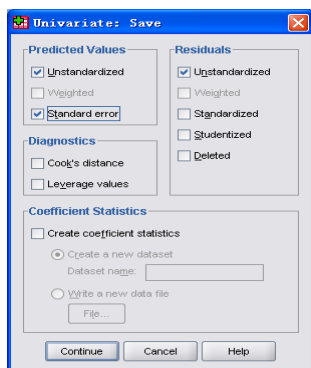


图 9-14 保存结果对话框

(1) Predicted Values 栏, 系统对每个观测量给出根据模型计算的预测值。

① Unstandardized, 给出非标准化预测值。

② Weighted, 如果在主对话框中选择了 WLS (Weighted Least Squares) 变量, 选中该复选项, 将保存加权的非标准化预测值。

③ Standard error, 给出预测值标准误。

(2) Diagnostics 诊断栏, 测量并标识对模型影响较大的观测量或自变量。

① Cook's distance, 给出 Cook 距离。

② Leverage values, 给出非中心化杠杆值。

#### (3) Residuals 残差栏

① Unstandardized, 给出非标准化残差值, 即观测值与预测值之差。

② Weighted, 如果在主对话框中选择了 WLS (Weighted Least Squares) 变量, 选中该复选项, 将保存加权的非标准化残差。

③ Standardized, 给出标准化残差, 又称 Pearson 残差。

④ Studentized, 给出学生化残差。

⑤ Deleted, 给出剔除残差, 因变量值与校正预测值之差。

以上选项给出的有关回归的统计量含义请参考第 11 章的有关内容。

#### (4) Save to New File 栏

选中 Coefficient statistics, 将模型参数估计的方差—协方差矩阵保存到一个新文件

中。对因变量产生三行数据：一行是参数估计值，一行是与参数估计值相对应的显著性检验的  $t$  统计量，还有一行是残差自由度。所生成的新数据文件可以作为另外分析的输入数据文件。单击 File 按钮，打开相应的保存对话框，指定文件的保存位置和文件名。

## 6. 选择输出项

在主对话框中，单击 Options 按钮，展开 Univariate: Options 对话框，见图 9-15。

### (1) Estimated Marginal Means 估计的边际均值栏。

① 在 Factor(s) and Factor Interactions 框中列出了在 Model 对话框中所指定的效应项。在该框中选定因素变量的各种效应项，单击移动箭头，将其复制到 Display Means for 框中。选择主效应，则产生估计的边际均值表。选择二维交互效应产生的估计边际均值表实际上是典型的单元格均值表。选择三维交互效应也显示单元格均值表。选择 OVERALL 项产生边际均值的均值。详见 2×2 析因方差分析例题。

② 在 Display Means for 框中有主效应时激活此框下面的 Compare main effects 复选项，对主效应的边际均值进行组间的配对比较。

③ Confidence interval adjustment 参数框列出了进行多重组间比较时置信区间和显著性水平调整方法的选项，打开下拉菜单，共有三个选项。

- LSD(none)，不进行调整。
- Bonferroni，邦弗伦尼方法，是基于 Student  $t$  统计量的方法。适用于要进行比较的均值，对数比较少的情況。
- Sidak，计算  $t$  统计量进行多重配对比较，调整多重比较的显著性水平。限制比 Bonferroni 检验更严格。

### (2) Display 栏，指定要输出的统计量。

① Descriptive statistics，输出的描述统计量有观测量均值、标准差和各单元格中的观测量数。

② Estimates of effect size，输出效应量估计。给出  $\eta^2$  (eta square)，它反映了每个效应与每个参数估计值可以归于因素的总变异的大小。

③ Observed power，给出各种检验假设的功效。计算功效的显著性水平，系统默认的临界值是 0.05。

④ Parameter estimates，给出各因素变量的模型参数估计、标准误、T 检验的  $t$  值、显著性概率和 95% 的置信区间。

⑤ Contrast coefficient matrix，显示变换系数矩阵或 L 矩阵。

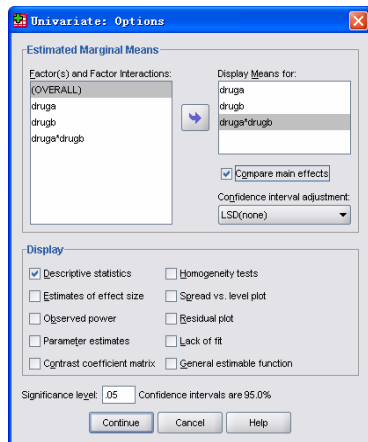


图 9-15 选择输出项的对话框



- ⑥ Homogeneity tests，进行方差齐性检验。
  - ⑦ Spread vs. level plot，绘制观测值均值—标准差图、观测值均值—方差图。
  - ⑧ Residuals plot，绘制残差图。给出观测值、预测值散点图和观测值数目对标准化残差的散点图，加上正态和标准化残差的正态概率图。
  - ⑨ Lack of fit，检查独立变量和非独立变量间的关系是否被充分描述。
  - ⑩ General estimable function，可以根据一般估计函数自定义假设检验。对比系数矩阵的行与一般估计函数是线性组合的。
- (3) 在 Significance level 框中改变 Confidence intervals 框内多重比较的显著性水平。

9.3.3 随机区组设计的方差分析实例

【例 4】四个种系未成年雌性大白鼠各三只，每只按一种剂量注射雌激素，一段时间后，解剖称子宫重量。数据见表 9-17，数据录入格式见表 9-16，数据编号 data09-03。

1. 操作方法与步骤

(1) 定义三个变量建立数据文件：两个分类变量，一个尺度（连续）变量。

表 9-17 不同种系、剂量的子宫重量

种系	剂 量		
	0.2 (1)	0.4 (2)	0.8 (3)
A (1)	106	116	145
B (2)	42	68	115
C (3)	70	111	133
D (4)	42	63	87

	mouse	etrogen	wuteri
1	1	1	106
2	1	2	116
3	1	3	145
4	2	1	42
5	2	2	68
6	2	3	115
7	3	1	70
8	3	2	111
9	3	3	133
10	4	1	42
11	4	2	63
12	4	3	87

图 9-16 方差分析的数据安排

大白鼠种系变量 mouse，取值 1~4，是种系 A~D 的代码。  
雌激素剂量变量 etrogen，取值 1~3，是剂量 0.2、0.4、0.8 三种剂量的代码。  
子宫重量变量 wuteri，连续变量。是本课题的研究对象。  
输入数据时应该注意观测值是如何构成的。正确的构成方式应该如图 9-16 所示。

(2) 按 Analyze→General Liner Model→Univariate 顺序单击菜单项，展开 Univariate 主对话框，如图 9-8 所示。

(3) 定义因变量和因素变量

- ① 定义 wuteri 为因变量：在源变量表中，选择 wuteri 变量进入 Dependent 框。
  - ② 定义 mouse 和 etrogen 变量为固定因素变量，选择并送入 Fixed Factor(s)框。
- (4) 单击 Model 按钮，展开 Model 对话框，选择自定义 Custom。
- ① 在 Building Terms 栏内的参数框中选择 Main effect 项，定义主效应。
  - ② 从 Factors & Covariates 框中分别选定 mouse，etrogen 并移入 Model 框中。

(5) 单击 OK 按钮, 执行多元方差分析过程。

执行的程序和结果如下:

UNIANOVA

wuteri BY mouse etrogen

/METHOD = SSPTYPE (3)

/INTERCEPT = INCLUDE

/CRITERIA = ALPHA (.05)

/DESIGN = mouse etrogen

①

②

③

④

⑤

⑥

程序语句解释

①调用 UNIANOVA 过程。②定义因变量和因素变量。③使用系统默认的分解偏差平方和的分解方法, 此为系统默认 Type III方式。④模型包括截距。⑤指定显著性水平。⑥指定随机区组设计的因素变量是 mouse, etrogen。

(6) 输出结果见表 9-18 和表 9-19。

表 9-18 因素变量表

Between-Subjects Factors			
		Value Label	N
大白鼠种系	1	A	3
	2	B	3
	3	C	3
	4	D	3
雌激素剂量	1	0.2	4
	2	0.4	4
	3	0.8	4

表 9-19 主效应方差分析检验结果

Tests of Between-Subjects Effects					
Dependent Variable: 子宫重量					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12531.667 <sup>a</sup>	5	2506.333	27.677	.000
Intercept	100467.000	1	100467.000	1109.452	.000
mouse	6457.667	3	2152.556	23.771	.001
etrogen	6074.000	2	3037.000	33.537	.001
Error	543.333	6	90.556		
Total	113542.000	12			
Corrected Total	13075.000	11			

a. R Squared = .958 (Adjusted R Squared = .924)

表 9-18 为变量信息, 大白鼠子宫重量按大白鼠种系和雌激素剂量分组。因素变量有种系 mouse, 取值 1~4, 是种系 A~D 的代码, 雌激素剂量 etrogen, 取值 1~3, 是剂量 0.2、0.4、0.8 的代码。N 是每一单元的样本含量。

表 9-19 是方差分析表, 在表的左上方标明研究的对象即因变量是 wuteri。

① Source 列: 偏差的来源。这一列表明此列右面将按以下各项列出各统计量:

- Corrected Model, 校正模型的偏差平方和。即经均值校正后的偏差平方和。我们在 Model 对话框中设置的方差分析模型只有两个主效应。该值等于两个主效应 mouse、etrogen 偏差平方和之和。

- Intercept, 截距的偏差平方和。

- 主效应 mouse, 其偏差平方和表明的是由于大白鼠种系不同(对雌激素反应不同)造成的子宫重量之差异, 与 etrogen 偏差平方和一样, 均属于组间偏差平方和。

- 主效应 etrogen, 其偏差平方和解释的是不同雌激素剂量造成的子宫重量之差异。

- Error, 误差。它是除去模型中指定的效应外不可解释的部分。一般情况下, 可能包括未考虑到的协变量效应或交互效应、随机因素效应和组内差异。在本问题中, 其偏

差平方和反映组内（即个体之间的）差异，也称组内偏差平方和。误差项用于检验各效应的假设。其均方值作为  $F$  检验计算  $F$  值的分母。

- **Total**，是因变量的总偏差平方和在数值上等于截距、两个主效应和误差的偏差平方和之总和。反映因变量原始的总变异。

- **Corrected Total**，校正的总偏差平方和。

对方差模型来说，从其值等于校正模型偏差平方和与误差之偏差平方和之总和可以看出，方差模型的总偏差平方和，分解为两个主效应（组间）偏差平方和与误差（组内）偏差平方和。

对于以 **wuteri** 为因变量、**mouse**、**etrogen** 为自变量的线性回归模型来说，**Corrected Total** 就是线性模型的总偏差平方和，在数值上等于回归平方和与残差平方和之总和。

② **Type III Sum of Squares** 列：偏差平方和。

③ **df** 列：自由度。

④ **Mean Square** 列：均方，在数值上等于偏差平方和除以相应的自由度。

⑤ **F** 列：即  $F$  值，是各效应项的均方与误差项的均方之比。

⑥ **Sig** 列：是进行  $F$  检验的  $p$  值。

从两个主效应的  $F$  检验结果的  $p$  值看， $p \leq 0.05$ ，由此得出种系 **moues** 和剂量 **etrogen** 对因变量 **wuteri** 在 0.05 水平上是有显著性差异的。截距的检验结果  $p < 0.05$ ，结论是：

对相同剂量的雌激素，不同种系大白鼠子宫重量增加明显不同。因数据较少，可从原始数据观察。

对同种系大白鼠，随雌激素剂量增加，子宫重量增加， $p \leq 0.05$  均值差异显著。

在 **wuteri** 因变量与 **mouse**、**etrogen** 两个自变量之间存在线性回归关系。

#### (7) 应注意的问题

本例中虽然有两个因素变量，但是两个因素变量的各水平构成的每个组合只有一个观测量。实际上这种实验设计只符合单因素方差分析的实验设计方案。因此如果分析因素间的交互作用，无法计算差异的显著性，因此输出结果不能给出  $F$  值及其概率。如果本例按照双因素设计进行方差分析，不考虑交互作用会得出较满意的结果。这就需要注意一定要使用 **Model** 选项，在 **Building Terms** 里只选择 **Main Effect** 项，而不要选 **Interaction** 项，即不要指定交互项。

### 9.3.4 $2 \times 2$ 析因实验方差分析实例

**【例 5】**本例为使用两种药物 A 和 B 治疗缺铁性贫血病人的数据，是一个  $2 \times 2$  析因实验设计的例题，主要说明均值对比的选项与结果。研究 A、B 两种治疗缺铁性贫血药物的疗效，随机选取 12 个病人分为 4 组，给以不同的治疗：第一组使用一般疗法；第二组使用一般疗法外加药物 A；第三组使用一般疗法外加药物 B；第四组使用一般疗法外加用药物 A 和药物 B。一个月后观察红细胞增加数（百万/ $\text{mm}^3$ ），作析因分析。数据编

号为 data09-04。

### 1. 数据说明与假设

因素变量两个: drugA 和 drugB, 两个变量均有两水平, 0 表示不用此药, 标签为“no”  
1 表示使用此药, 标签为“yes”。因变量: redcell (红细胞增加数), 单位: 百万/mm<sup>3</sup>。

该研究的检验假设是  $H_0$ : 药物 A 和药物 B 对患者红细胞增加无显著效果。两种药物无协同作用 (即无交互效应)。

### 2. 操作步骤

(1) 读取数据 data09-04。按 Analyze→General Liner Model→Univariate 顺序单击菜单项, 最后打开 Univariate 主对话框。

(2) 指定分析变量: 将变量 redcell 移入 Dependent 框, 为因变量。将 drugA 和 drugB 变量进入 Fixed Factor(s)框, 作为因素变量。

(3) 由于本次分析为全模型, 因此不用对 Model 对话框作任何操作。全模型即模型中包括所有主效应和交互效应。对于双因素的全模型应该包括两个主效应 drugA、drugB, 一个交互效应 drugA\*drugB。

(4) 在主对话框中, 单击 Plots 按钮, 展开相应的对话框, 要求作三个图的操作如下:

① 在 Factors 框中选择 drugA, 送入横坐标栏。单击 Add 按钮, 在 Plots 栏中出现图形表达式 drugA。

② 在 Factors 框中选择 drugB, 送入横坐标栏。单击 Add 按钮, 在 Plots 栏中出现图形表达式 drugB。

③ 在 Factors 框中选择 drugA, 送入横坐标栏, 在 Factors 框中选择 drugB, 送入分线栏。单击 Add 按钮, 在 Plot 栏中出现图形表达式 drugA\*drugB。

④ 在 Display 栏内选择 Descriptive statistics, 见图 9-15。

⑤ 单击 Continue 按钮, 返回主对话框。

(5) 主对话框中, 单击 Options 按钮, 展开相应的对话框。在 Factors 框中分别选择因素变量 drugA、drugB、drugA\*drugB 和 Overall 单击向右箭头按钮将因素变量送入 Display Means for 框中。单击 Continue 回到主对话框。

(6) 单击 OK 按钮, 执行命令。

### 3. 命令程序与解释。

#### (1) 程序清单

UNIANOVA

redcell BY drugA drugB ①

/METHOD=SSTYPE (3) ②

/INTERCEPT=INCLUDE ③

/PLOT=PROFILE (druga drugb drugA\*drugB) ④

/EMMEANS=TABLES (drugA) ⑤

```
/EMMEANS=TABLES (drugB)
/EMMEANS=TABLES (drugA*drugB)
/EMMEANS = TABLES (OVERALL)
/PRINT=DESCRIPTIVE
/CRITERIA=ALPHA (.05)
/DESIGN=drugA drugB drugA*drugB.
```

(2) 命令程序解释

① UNIANOVA 命令指定因变量 redcell，BY 后面是两个因素变量 drugA、drugB。

② METHOD 子命令指定平方和分解方法是系统默认的 Type III。

③ INTERCEPT 子命令指定模型中包括截距。

④ PLOT 子命令指定做边际图，纵坐标是因变量，共 3 个图：

- 横轴变量是 drugA。
- 横轴变量是 drugB。
- 前面一个变量是横轴变量 drugA，后面的是分线变量，是 drugB。

⑤~⑦ EMMEANS 子命令指定分别输出两个因素变量和两个因素变量交互项的边际均值表，及综合边际均值表。

⑧ PRINT 子命令指定输出描述统计量。

⑨ CRITERIA 子命令指定  $\alpha=0.05$ 。

⑩ DESIGN 子命令指定分析两个因素的主效应和交互效应全模型。

4. 输出结果见表 9-20 至表 9-26、图 9-17 至图 9-19。

5. 结果说明与分析

表 9-20 是“两种药物对红细胞增加数作用的研究”课题中的变量信息。表中列出 drugA 和 drugB 两个因素变量和分类水平，以及每个水平的样本含量。

表 9-21 为描述统计量。

表 9-22 为方差分析结果。可以看出：

- 总校正偏差平方和分解为校正模型的和随机误差的偏差平方和。校正模型的偏差平方和=drugA 偏差平方和+drugB 偏差平方和+交互效应 drugA\*drugB 的偏差平方和。
- 随机误差偏差平方和为 0.08。
- 各项偏差平方和除以各自的自由度是相应的均方。各项  $F$  值为各项均方除以误差均方。
- $F$  检验的结果，显著性概率  $p$  值均小于 0.01。

表 9-20 因素变量表

Between-Subjects Factors			
		Value Label	N
drugA	0	no	6
	1	yes	6
drugB	0	no	6
	1	yes	6

表 9-21 描述统计量

Descriptive Statistics				
Dependent Variable: red cell				
drug A	drugB	Mean	Std. Deviation	N
no	no	.800	.1000	3
	yes	1.000	.1000	3
	Total	.900	.1414	6
yes	no	1.200	.1000	3
	yes	2.100	.1000	3
	Total	1.650	.5010	6
Total	no	1.000	.2366	6
	yes	1.550	.6091	6
	Total	1.275	.5259	12

表 9-22 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: red cell					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2.963 <sup>a</sup>	3	.988	98.750	.000
Intercept	19.508	1	19.508	1950.75	.000
druga	1.688	1	1.688	168.750	.000
drugb	.908	1	.908	90.750	.000
druga * drugb	.368	1	.368	36.750	.000
Error	.080	8	.010		
Total	22.550	12			
Corrected Total	3.043	11			

a. R Squared = .974 (Adjusted R Squared = .964)

结论: drugA、drugB 均对红细胞的增加有显著疗效。并且交互效应也很显著。检验结果拒绝无效假设, 使用药物 A 与不使用药物 A 的红细胞增加数的均值有显著性差异。使用药物 B 与不使用药物 B 的红细胞增加数的均值有显著性差异。同时使用药物 A 和药物 B, 两药物协同作用也很显著。

表 9-23 至表 9-26 为红细胞增加数的估计的边际值表。总结这 4 个表成表 9-27。可以看出交互项产生单元格中均数, 主效应项生成边际均数, 总效应项 OVERALL 生成总均数 1.275。

表9-23 drugA边际均值估计值表

1. drugA				
Dependent Variable: red cell				
drugA	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
no	.900	.041	.806	.994
yes	1.650	.041	1.556	1.744

表9-24 drugB边际均值估计值表

2. drugB				
Dependent Variable: red cell				
drugB	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
no	1.000	.041	.906	1.094
yes	1.550	.041	1.456	1.644

表9-25 交互项边际均值估计值表

3. drugA * drugB				
Dependent Variable: red cell				
drugA	drugB	Mean	Std. Error	95% Confidence Interval
				Lower Bound Upper Bound
no	no	.800	.058	.667 .933
	yes	1.000	.058	.867 1.133
yes	no	1.200	.058	1.067 1.333
	yes	2.100	.058	1.967 2.233

表9-26 综合边际均值估计值表

4. Grand Mean			
Dependent Variable: red cell			
Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
1.275	.029	1.208	1.342

图 9-17、图 9-18 和图 9-19 是一系列边际均值图。读者可以对照图 9-27 查看数据与图的关系。从表 9-17 和表 9-18 可以看出, 每种药物单独效应: 用药与不用药对红细胞增加数效用。在表 9-19 中可以看出两直线明显不平行, 因此很明显, 这两种药之间存在交互效应。

如果想更直观地比较 4 种疗法的疗效, 可以根据 drugA、drugB 使用 Transform 菜单中的第一项 Compute 功能生成有 4 水平的新变量, 分别代表一般治疗、一般治疗加 A 药、一般治疗加 B 药和一般治疗加 A 药和 B 药。利用多重比较功能比较 4 种用药方法的疗效。

表 9-27 边际值估计值示意图

实验分组		A 药		B 边际均值
		不用	使用	
B 药	不用	0.80	1.20	1.000
	使用	1.00	2.10	1.550
A 边际均值		0.900	1.650	1.275

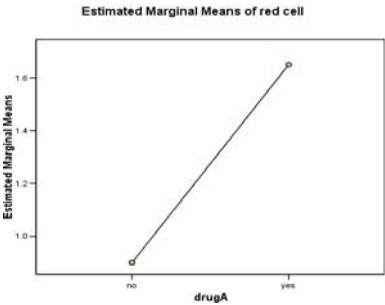


图 9-17 A 药效应红细胞增加数均值图

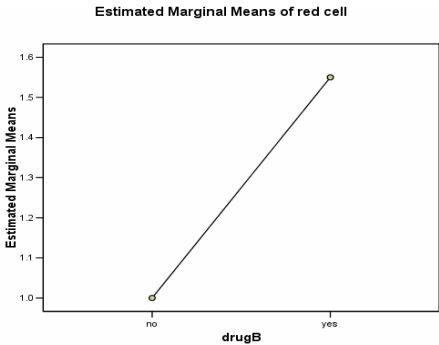


图 9-18 B 药效应红细胞增加数均值图

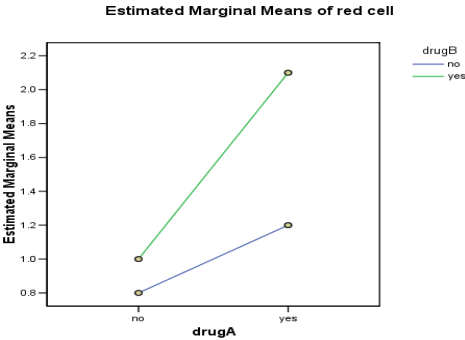


图 9-19 A、B 药对红细胞增加数交互效应边际图

9.3.5 拉丁方区组设计的方差分析实例

【例 6】拉丁方实验设计的特点是有两个以上因素变量，每个因素变量的水平数相等。

变量: variety (甜菜种系)、rep (地块行)、col (地块列) havrvest (收获次数)、yield (产量)。要求分析六种甜菜品种在相同土壤条件下的产量是否有显著性差异。为了得出这一结论，同时检验地块是否对平均产量有影响。即地块的行与行之间、列与列之间的平均产量是否有显著性差异。将六种甜菜种子播在六行、六列的地块上，记录两次收获的产量。数据编号 data09-05。实验的假设是：不同地块（行、列）对产量均值无影响，不同种子产量均值间也无显著差异。

- 1. 操作步骤。分两步完成，先做方差分析，再做边际值估计值表。
  - (1) 读取数据 data09-05。按 Analyze→General Liner Model→Univariate 顺序单击菜单项，最后打开 Univariate 主对话框。
  - (2) 在主对话框中定义分析变量
    - ① 将 yield 变量移入 Dependent 框作为因变量。
    - ② 将 rep、col、variety 变量进入 Fixed Factor(s)框，这些变量作为因素变量。

(3) 在主对话框中，单击 **Model** 按钮，展开相应的对话框。在对话框中选择 **Custom**，自定义模型：指定要求分析三个主效应 **rep**、**col**、**variety**。选择结束后，单击 **Continue** 按钮，返回主对话框。

(4) 在主对话框中，单击 **Options** 按钮，展开相应的对话框。选择三个因素变量 **rep**、**col**、**variety** 和 **Overall** 送入 **Display Means for** 栏内，选择 **Compare main effects**。其他使用默认值。

(5) 单击 **OK** 按钮，执行一次，完成方差分析。

2. 输出结果见表 9-28 至表 9-32。

表 9-28 为方差分析表，只对 **rep**、**col**、**variety** 变量的主效应作方差分析。方差分析解决三个因素变量的各水平，产量平均值之间差异是否有统计意义。

表 9-28 方差分析表

Tests of Between-Subjects Effects					
Dependent Variable: YIELD					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	27.717 <sup>a</sup>	15	1.848	1.339	.211
Intercept	22588.751	1	22589	16364.072	.000
rep	4.460	5	.892	.646	.666
col	1.695	5	.339	.246	.940
variety	21.563	5	4.313	3.124	.015
Error	77.302	56	1.380		
Total	22693.770	72			
Corrected Total	105.019	71			

a. R Squared = .264 (Adjusted R Squared = .067)

表9-29 各列、各行、各种甜菜产量的分类均值表（边际均值）

Estimates					Estimates					Estimates				
Dependent Variable: YIELD					Dependent Variable: YIELD					Dependent Variable: YIELD				
REP	Mean	Std. Error	Lower Bound	Upper Bound	COL	Mean	Std. Error	Lower Bound	Upper Bound	VARIETY	Mean	Std. Error	Lower Bound	Upper Bound
1	17.850	.339	17.171	18.529	1	17.483	.339	16.804	18.163	1	17.367	.339	16.687	18.046
2	17.658	.339	16.979	18.338	2	17.650	.339	16.971	18.329	2	17.817	.339	17.137	18.496
3	18.017	.339	17.337	18.696	3	17.642	.339	16.962	18.321	3	17.475	.339	16.796	18.154
4	17.933	.339	17.254	18.613	4	17.942	.339	17.262	18.621	4	17.367	.339	16.687	18.046
5	17.517	.339	16.837	18.196	5	17.875	.339	17.196	18.554	5	18.883	.339	18.204	19.563
6	17.300	.339	16.621	17.979	6	17.683	.339	17.004	18.363	6	17.367	.339	16.687	18.046

查看各主效应的 **Sig of F** 值，只有因素变量 **variety** 的值为 0.015 小于 0.05。可得出结论：六种甜菜的平均产量具有显著性差异。平均产量的差异主要是品种不同造成的。

表 9-29 为各列、行和各个品种的边际均值估计值表，此外还有标准误和区间估计。

表 9-30 的三个表为每个因素的各水平均值的成对比较表。每个表中给出各变量两两水平之间的均值之差、均值差的标准误、针对两均值相等的假设检验的显著性概率 **Sig** 值、差值的 95% 置信区间。从三个表中可以看到只有第 5 种子比其他 5 种种子产量都高，且差值具有明显的统计意义。

表 9-31 的三个表为各因素单变量方差分析表。表中给出 **F** 值及大于 **F** 临界值的概率。可以看出只有种类的方差分析的 **Sig** 值为 0.015，小于 0.05。

综上所述，产量主要受种子的影响，而第 5 种种子的产量明显高于其他种子。产量与地块所处位置行、列无关。

表 9-32 是最后给出的综合统计表，给出产量的总均值、均值标准误和 95% 置信区间。



表9-30 主效应因素均值表（rep、col、variety）

Pairwise Comparisons

Dependent Variable: YIELD

(I) REP	(J) REP	Mean Difference (I-J)	Std. Error	Sig. <sup>a</sup>	Lower Bound	Upper Bound
1	2	-.192	.480	.691	-.769	1.153
3	3	-.167	.480	.730	-1.128	.794
4	4	-.083	.480	.863	-1.044	.878
5	5	.333	.480	.490	-.628	1.294
6	6	.550	.480	.256	-.411	1.511
2	1	-.192	.480	.691	-1.153	.769
3	3	-.358	.480	.458	-1.319	.603
4	4	-.275	.480	.569	-1.236	.686
5	5	-.142	.480	.769	-.819	1.103
6	6	.358	.480	.458	-.603	1.319
3	1	.167	.480	.730	-.794	1.128
2	3	.358	.480	.458	-.603	1.319
4	4	.083	.480	.863	-.878	1.044
5	5	.500	.480	.302	-.461	1.461
6	6	.717	.480	.141	-.244	1.678
4	1	.083	.480	.863	-.878	1.044
2	2	.275	.480	.569	-.686	1.236
3	3	-.083	.480	.863	-1.044	.878
5	5	.417	.480	.389	-.544	1.378
6	6	.633	.480	.192	-.328	1.594
5	1	-.333	.480	.490	-1.294	.628
2	2	-.142	.480	.769	-1.103	.819
3	3	-.500	.480	.302	-1.461	.461
4	4	-.417	.480	.389	-1.378	.544
6	6	.217	.480	.653	-.744	1.178
6	1	-.550	.480	.256	-1.511	.411
2	2	-.358	.480	.458	-1.319	.603
3	3	-.717	.480	.141	-1.678	.244
4	4	-.633	.480	.192	-1.594	.328
5	5	-.217	.480	.653	-1.178	.744

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Pairwise Comparisons

Dependent Variable: YIELD

(I) COL	(J) COL	Mean Difference (I-J)	Std. Error	Sig. <sup>a</sup>	Lower Bound	Upper Bound
1	2	-.167	.480	.730	-1.128	.794
3	3	-.158	.480	.743	-1.119	.803
4	4	-.458	.480	.343	-1.419	.503
5	5	-.392	.480	.418	-1.353	.569
6	6	-.200	.480	.678	-1.161	.761
2	1	.167	.480	.730	-.794	1.128
3	3	.008	.480	.986	-.953	.969
4	4	-.292	.480	.546	-1.253	.669
5	5	-.225	.480	.641	-1.186	.736
6	6	-.033	.480	.945	-.994	.928
3	1	.158	.480	.743	-.803	1.119
2	2	-.008	.480	.986	-.969	.953
4	4	-.300	.480	.534	-1.261	.661
5	5	-.233	.480	.629	-1.194	.728
6	6	-.042	.480	.931	-1.003	.919
4	1	.458	.480	.343	-.503	1.419
2	2	.292	.480	.546	-.669	1.253
3	3	.300	.480	.534	-.661	1.261
5	5	.067	.480	.890	-.894	1.028
6	6	.258	.480	.592	-.703	1.219
5	1	.392	.480	.418	-.569	1.353
2	2	.225	.480	.641	-.736	1.186
3	3	.233	.480	.629	-.728	1.194
4	4	-.067	.480	.890	-1.028	.894
6	6	.192	.480	.691	-.769	1.153
6	1	.200	.480	.678	-.761	1.161
2	2	.033	.480	.945	-.928	.994
3	3	.042	.480	.931	-.919	1.003
4	4	-.258	.480	.592	-1.219	.703
5	5	-.192	.480	.691	-1.153	.769

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Pairwise Comparisons

Dependent Variable: YIELD

(I) VAR IETY	(J) VAR IETY	Mean Difference (I-J)	Std. Error	Sig. <sup>a</sup>	Lower Bound	Upper Bound
1	2	-.450	.480	.352	-1.411	.511
3	3	-.108	.480	.822	-1.069	.853
4	4	2.33E-015	.480	1.000	-.961	.961
5	5	-.1517*	.480	.003	-2.478	-.556
6	6	2.71E-014	.480	1.000	-.961	.961
2	1	.450	.480	.352	-.511	1.411
3	3	.342	.480	.479	-.619	1.303
4	4	.450	.480	.352	-.511	1.411
5	5	-.1067*	.480	.030	-2.028	-.106
6	6	.450	.480	.352	-.511	1.411
3	1	.108	.480	.822	-.853	1.069
2	2	-.342	.480	.479	-1.303	.619
4	4	.108	.480	.822	-.853	1.069
5	5	-.1408*	.480	.005	-2.369	-.447
6	6	.108	.480	.822	-.853	1.069
4	1	-.23E-015	.480	1.000	-.961	.961
2	2	-.450	.480	.352	-1.411	.511
3	3	-.108	.480	.822	-1.069	.853
5	5	-.1517*	.480	.003	-2.478	-.556
6	6	2.48E-014	.480	1.000	-.961	.961
5	1	1.517*	.480	.003	.556	2.478
2	2	1.067*	.480	.030	.106	2.028
3	3	1.408*	.480	.005	.447	2.369
4	4	1.517*	.480	.003	.556	2.478
6	6	1.517*	.480	.003	.556	2.478
6	1	-2.7E-014	.480	1.000	-.961	.961
2	2	-.450	.480	.352	-1.411	.511
3	3	-.108	.480	.822	-1.069	.853
4	4	-2.5E-014	.480	1.000	-.961	.961
5	5	-.1517*	.480	.003	-2.478	-.556

Based on estimated marginal means

\*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-31 单变量方差分析的 3 个表（rep、col、variety）

Univariate Tests

Dependent Variable: YIELD

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	4.460	5	.892	.646	.666
Error	77.302	56	1.380		

The F tests the effect of REP. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

Univariate Tests

Dependent Variable: YIELD

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	1.695	5	.339	.246	.940
Error	77.302	56	1.380		

The F tests the effect of COL. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

Univariate Tests

Dependent Variable: YIELD

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	21.563	5	4.313	3.12	.015
Error	77.302	56	1.380		

The F tests the effect of VARIETY. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

表 9-32 总均值表

1. Grand Mean

Dependent Variable: YIELD

Mean	Std. Error	95% Confidence Interval
17.713	.138	Lower Bound 17.435 Upper Bound 17.990

本例中虽然有三个因素变量，但是三个因素变量的各水平组合构成的每个单元只有一个观测量。实际上这种实验设计如果分析因素间的交互作用，无法计算差异的显著性，因此输出结果不能给出  $F$  值及其概率。如果本例按照三因素设计进行方差分析，不考虑交互作用会得出较满意的结果。这就需要注意一定要使用 **Model** 选项，在 **Building Terms** 里只选择 **Main Effect** 项，而不要选择 **Interaction** 项，即不要指定交互项。所以我们在进

行方差分析时不考虑交互作用，而只考虑主效应，要求边际均值时才用全模型。

### 9.3.6 协方差分析实例

协方差分析是利用线性回归方法消除混杂因素的影响后进行的方差分析。就是说先从因变量的总偏差平方和中去掉协变量对因变量的回归平方和，再对残差平方和进行分解，进行方差分析。例如考虑药物对患者某个生化指标变化的影响，要比较实验组与对照组该指标的变化均值是否有显著性差异以确定药物的有效性，可能要考虑患者病程的长短、年龄以及原指标水平对疗效的影响。要消除这些因素的影响，考虑药物疗效，才是科学的分析方法。有些实验可以考虑观测对象的选择，使这些条件都一致。例如选择同品种、同一胎的大白鼠分组，在相同的饲养条件下进行实验，可以相应地避免许多混杂因素的影响。其他实验很难避免，因此要考虑使用协方差分析方法。这些混杂因素变量称作协变量。

协方差分析中要求因变量应该是等间隔测量的变量（Scale 型），理论上要求正态分布。因素变量是分类变量，并且相互独立。协变量是与因变量存在一定相关关系（相互不独立）的等间隔测量的变量。因变量与协变量之间是否线性相关，可以通过经验得知或使用 Graphs 菜单中的 scatter 命令作散点图来做初步的直观判断。

【例 7】数据文件 data09-06 是镉作业工人年龄与肺活量的数据，数据来源于《医用统计方法》（金丕焕，人民卫生出版社）。镉作业工人按暴露于镉粉尘的年数分为大于等于 10 年和不足 10 年两组。两组工人的年龄未经控制（人随着年龄的增长，肺活量也会有所下降），测量了每个工人的肺活量。课题研究暴露于镉粉尘中的年数与肺活量的关系。数据变量如下：

time（接触镉粉尘时间分组），取值 1 代表大于等于 10 年、2 代表不足 10 年；age（年龄）；vitalcp（肺活量，单位：升）。

#### 1. 操作步骤

(1) 读取数据 data09-06。按 Analyze→General Liner Model→Univariate 顺序单击菜单项，打开 Univariate 主对话框。

(2) 指定分析变量：将 vitalcp 变量移入 Dependent Variable 框，即 vitalcp 变量为因变量；将 time 变量送入 Fixed Factor(s)框，即 time 变量作为因素变量；将 age 变量送入 Covariate(s)框，即 age 变量作为协变量。

(3) 在主对话框中，单击 Options 按钮，展开相应的对话框。

① Factor(s) and Factor Interactions 框中选择因素变量 time，将其送入 Display Means for 框。要求输出暴露于镉粉尘年数大于等于 10 年、不足 10 年两组工人的肺活量平均值。

② 在 Display 栏内选中 Parameter estimates，要求输出年龄做自变量，肺活量做因变量的线性回归方程的参数。

(4) 单击 Continue 按钮，返回主对话框。单击 OK 按钮，提交系统执行。

2. 程序清单与程序说明

(1) 程序清单：

UNIANOVA

Vitalcp BY time WITH age

/METHOD=SSSTYPE(3)

/INTERCEPT=INCLUDE

/EMMEANS=TABLES(time) WITH(age=MEAN)

/PRINT=PARAMETER

/CRITERIA=ALPHA(.05)

/DESIGN=age time.

①

②

③

④

⑤

⑥

⑦

⑧

(2) 程序说明

① 调用 UNIANOVA 过程。

② 指定因变量为肺活量 vitalcp，因素变量 time，协变量 age。BY time 指定按变量 time 分组进行均值差异显著性分析。可称 time 为方差分析的 BY 变量，即因素变量。WITH age 指定变量 age 是协变量。也称为 WITH 变量。

③ 用系统默认的 Type III 方法计算离差平方和。

④～⑧分别为指定模型中包括截距；指定按 time 分组求平均值；指定打印输出参数估计值；显著性水平，设定临界值 $\alpha=0.05$ 。设计分析 time 的主效应和 age 的协变量效应。

3. 输出结果见表 9-33 至表 9-36。

表 9-33 为因素变量表。列出了按时间分组的变量标签、样本量。

表 9-34 为方差分析结果：

表9-34 方差分析表

表 9-33 因素变量表

Between-Subjects Factors			
	Value Label	N	
exposed 1	>=10years	12	
time 2	<10years	16	

Tests of Between-Subjects Effects

Dependent Variable: vital capacity

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	11.085 <sup>a</sup>	2	5.543	10.073	.001
Intercept	41.936	1	41.936	76.216	.000
age	10.881	1	10.881	19.775	.000
time	.542	1	.542	.985	.330
Error	13.755	25	.550		
Total	483.625	28			
Corrected Total	24.841	27			

<sup>a</sup>. R Squared = .446 (Adjusted R Squared = .402)

① 表的左上方列出因变量（Dependent Variable）为 vitalcp，即研究对象为肺活量。

② 分别列出方差来源，系统默认的 TYPE III 的偏差平方和，自由度，均方差，F 值和 Sig。

③ 从方差分析表中看到总的偏差平方和 24.841 被分解为条件引起的平方和（Corrected Model）11.085 和实验误差引起的平方和（Error）13.755。从显著性概率（Sig）看，time 的概率 0.330，大于 0.05。协变量效应由协变量 age 决定，其偏差平方和为 10.881， $p=0.000$ ，小于 0.001。因此可以得出结论，肺活量的差异是由于被试者的年龄差异所致，

与被试者接触镉粉尘的时间是否大于 10 年无关。

表 9-35 是在 Options 对话框中的 Display 栏内选中了 Parameter estimates 的输出结果。这里主要给出了 age 作为自变量, vitalcp 作为因变量的线性回归方程的斜率, 即变量 age 的回归系数值为 -0.087。这一回归系数也是符合生理常识的。因为成年人随着年龄的增长, 肺活量会有所下降。

表 9-36 是在 Options 对话框中, 将 time 移入 Display Means for 框的结果。表中按 time 分组分别列出平均值、标准误和 95% 的置信区间。因素变量各单元均值是 10 年以下组的肺活量均值为 4.219, 10 年以上组的肺活量均值为 3.919。协方差分析结果表明, 这两组肺活量均值差异无统计意义上的显著性。

表 9-35 参数估测值的输出结果表

Parameter Estimates						
Dependent Variable: vital capacity						
Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	7.977	.886	8.998	.000	6.151	9.803
age	-.087	.020	-4.447	.000	-.127	-.047
[time=1]	.300	.303	.993	.330	-.323	.924
[time=2]	0 <sup>a</sup>	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

表 9-36 按时间分组的肺活量均值表

exposed time				
Dependent Variable: vital capacity				
exposed time	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
>=10years	4.219 <sup>a</sup>	.223	3.761	4.678
<10years	3.919 <sup>a</sup>	.191	3.526	4.312

a. Covariates appearing in the model are evaluated at the following values: age = 46.64.

### 9.3.7 多维交互效应方差分析实例

【例 8】例题主要表明使用 Univariate 过程进行多因素方差分析构成模型的灵活性。

1. 实验数据为教育心理学实验, 心理运动测验分数与被试者必须瞄准的目标大小关系的资料。

- (1) 选择四个大小不同的目标 target: 1 (T1), 2 (T2), 3 (T3), 4 (T4)。
- (2) 从若干使用过的设备中选择三部测验设备 device: 1 (D1), 2 (D2), 3 (D3)。
- (3) 选择两种不同明暗程度的照明环境 light: 1 (L1), 2 (L2)。

四个大小不同的目标、三部设备、两种不同的照明环境构成 4×3×2 的析因实验设计。不同目标、设备与照明水平构成了 24 个组合的单元。每一个组合中随机部署 5 名被试者进行测试心理运动得分。得到 120 个得分数据。

每个观测量为被试者在同一条件组合下的 5 个得分。数据编号为 data09-07。

#### 2. 操作步骤

(1) 按 Analyze→General Liner Model→Univariate 顺序单击菜单项, 打开 Univariate 主对话框。

(2) 将 score 变量移入 Dependent Variable 框, 测试得分为因变量; 将 target、device、light 变量进入 Fixed Factor(s)框, 作为因素变量。

(3) 在主对话框中, 单击 Model 按钮, 展开相应的对话框。首先选择 Custom 项, 自定义模型, 激活对话框中的各控制功能。

① 选择主效应。在 Build Term(s)栏中的参数框内选择 Main effects 主效应项。在 Factor(s) and Factor 框中选择 target、device、light 变量进入 Model 框中, 即这三个作为主效应定义到模型中。

② 选择交互项。在 Build Term(s)栏中的参数框内选择 Interaction 交互效应项。在 Factor & Covariates 框中, 选择一个变量 target, 按住 Ctrl 键, 选择第 2 个变量 device, 按一次箭头按钮, 将 target\*device 一个交互项移入到 Model 框中, 即该交互项进入模型。再用同样方法在模型中建立另一个二次交互项 Target\*light。

与建立二次交互项同样的方法在模型中建立三次交互项 Target\*device\*light。

(4) 在主对话框中, 单击 Options 按钮, 展开选项对话框。在 Factor(s) and Factor Interactions 框中指定 target\*device\*light, 将其送入 Display Means for 框, 目的是要输出各单元格的均值。

(5) 在主对话框中, 单击 Plot 按钮, 打开相应对话框, 选择 target 变量做横轴变量, 选择 device 变量作为分线变量, 选择 light 变量做分图变量, 分别送入右边的三个栏中。然后, 单击 Add 按钮, 将做图表达式 target\*device\*light 送入 Plots 框中。

(6) 单击 OK 按钮, 执行运算。

### 3. 生成的程序与程序解释

#### (1) 命令程序清单

UNIANOVA	①
score BY target device light	②
/METHOD=SSPSTYPE(3)	③
/INTERCEPT=INCLUDE	④
/PLOT = PROFILE( target*device*light )	⑤
/EMMEANS=TABLE(device*light*target)	⑥
/CRITERIA=ALPHA(.05).	⑦
/DESIGN target device light device*target light*target device*light*target	⑧

#### (2) 命令程序解释

- ① 调用 UNIANOVA 过程。
- ② 命令指定因变量为 score, BY 后面是三个因素变量。
- ③ 选择分解平方和方法是系统默认的 Type III。
- ④ 指定回归模型包含截距。
- ⑤ 指定以 target 变量为横轴、device 为分线变量、light 为分图变量作边际均值图。
- ⑥ 要求计算所有单元格的均值。
- ⑦ 选择的显著性水平临界值是  $\alpha=0.05$ 。
- ⑧ ANALYSIS 子命令指定分析模型。

因变量 score 方差源按三个因素变量主效应, 二维交互效应 target\*light、target\*device,

和三维交互效应来分析。

4. 运行结果见表 9-37 至表 9-39 和图 9-21、图 9-22。

表 9-37 为原始数据综合信息：系统接受了 120 个观测量；列出各个因素变量，变量值标签和样本含量。

表 9-38 为方差分析表，表的左上方标有因变量 score。

表 9-37 因素变量表

Between-Subjects Factors		
	Value Label	N
target	t1	30
	t2	30
	t3	30
	t4	30
device	d1	40
	d2	40
	d3	40
light	l1	60
	l2	60

表 9-38 方差分析结果

Tests of Between-Subjects Effects					
Dependent Variable: score					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	783.467 <sup>a</sup>	23	34.064	46.451	.000
Intercept	3162.133	1	3162.133	4312.000	.000
target	235.200	3	78.400	106.909	.000
device	86.467	2	43.233	58.955	.000
light	76.800	1	76.800	104.727	.000
target * device	104.200	6	17.367	23.682	.000
target * light	93.867	3	31.289	42.667	.000
target * device * light	186.933	8	23.367	31.864	.000
Error	70.400	96	.733		
Total	4016.000	120			
Corrected Total	853.867	119			

a. R Squared = .918 (Adjusted R Squared = .898)

表 9-39 各单元格观测测量均值

device * light * target						
Dependent Variable: score						
device	light	target	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
d1	l1	t1	2.000	.383	1.240	2.760
		t2	8.000	.383	7.240	8.760
		t3	8.800	.383	8.040	9.560
		t4	7.200	.383	6.440	7.960
	l2	t1	1.600	.383	.840	2.360
		t2	4.400	.383	3.640	5.160
		t3	5.600	.383	4.840	6.360
		t4	6.800	.383	6.040	7.560
d2	l1	t1	.800	.383	.040	1.560
		t2	8.400	.383	7.640	9.160
		t3	9.600	.383	8.840	10.360
		t4	9.200	.383	8.440	9.960
	l2	t1	5.200	.383	4.440	5.960
		t2	6.400	.383	5.640	7.160
		t3	4.800	.383	4.040	5.560
		t4	2.800	.383	2.040	3.560
d3	l1	t1	4.000	.383	3.240	4.760
		t2	5.200	.383	4.440	5.960
		t3	6.000	.383	5.240	6.760
		t4	2.000	.383	1.240	2.760
	l2	t1	3.200	.383	2.440	3.960
		t2	1.200	.383	.440	1.960
		t3	4.400	.383	3.640	5.160
		t4	5.600	.383	4.840	6.360

device * light * target						
Dependent Variable: score						
light	device	target	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
l1	d1	t1	2.000	.383	1.240	2.760
		t2	8.000	.383	7.240	8.760
		t3	8.800	.383	8.040	9.560
		t4	7.200	.383	6.440	7.960
	d2	t1	.800	.383	.040	1.560
		t2	8.400	.383	7.640	9.160
		t3	9.600	.383	8.840	10.360
		t4	9.200	.383	8.440	9.960
	d3	t1	4.000	.383	3.240	4.760
		t2	5.200	.383	4.440	5.960
		t3	6.000	.383	5.240	6.760
		t4	2.000	.383	1.240	2.760
l2	d1	t1	1.600	.383	.840	2.360
		t2	4.400	.383	3.640	5.160
		t3	5.600	.383	4.840	6.360
		t4	6.800	.383	6.040	7.560
	d2	t1	5.200	.383	4.440	5.960
		t2	6.400	.383	5.640	7.160
		t3	4.800	.383	4.040	5.560
		t4	2.800	.383	2.040	3.560
	d3	t1	3.200	.383	2.440	3.960
		t2	1.200	.383	.440	1.960
		t3	4.400	.383	3.640	5.160
		t4	5.600	.383	4.840	6.360

表 9-39 列出了三个因素变量构成的单元格表，给出了各单元格的均值、标准误和 95% 的置信区间。左表与右表完全一致，只是使用 Pivot Control 将表格中变量位置进行了调换。从均值列数据可以看出，light=1、device=2、target=3 这个条件组合的测试平均

分值最高。light=1、devise=2、target=1 的条件组合的测试平均分最低。心理学专业人士可以根据均值表和方差分析表得出专业性的结论。

图 9-20 和图 9-21 更直观地表现了表 9-38 中 Mean 栏中的均值数据。而且可以很清楚地看出在不同的光照条件下目标变量与设备之间均存在交互效应。读者可以自己作出使用不同设备时，光照与目标之间的边际均值图。

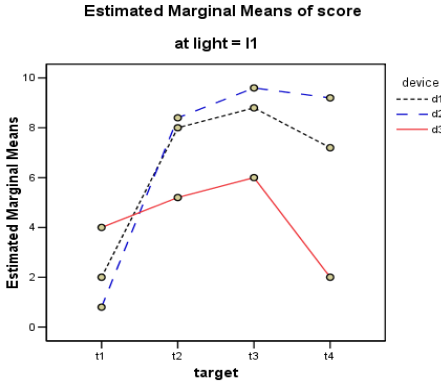


图 9-20 第一种照度下心理得分边际均值图

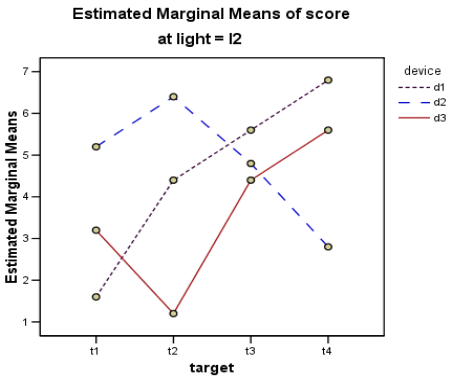


图 9-21 第二种照度下心理得分边际均值图

## 9.4 多因变量线性模型的方差分析

### 9.4.1 多因变量方差分析概述

GLM Multivariate 过程提供多因变量的方差分析。多因变量方差分析模型的因变量是尺度变量（连续变量）。分类变量作为固定因素变量，协变量必须是尺度变量。该模型是基于尺度因变量与作为预测因子的因素变量和协变量之间的相关关系。GLM Multivariate 过程构造的模型是一般线性模型。可以检验因变量在因素变量各水平组合中的组均值的效应，可以研究因素间的交互效应和单一因素的效应，另外还包括协变量效应和协变量与因素间的交互效应。对回归分析，协变量作为自变量即预测变量。

GLM Multivariate 过程可以检验平衡和不平衡模型。模型中每个单元包括相同数量的观测量为平衡设计。在多因变量模型中，模型中的效应平方和和误差平方和是矩阵形式的，而不是像在单因变量模型中的格式，这些矩阵称作 SSCP 矩阵（平方和和叉积矩阵）如果指定了不止一个因变量，多因素方差分析使用 Pillai 迹、Wilks  $\lambda$ 、Hotelling 迹和 Roy 最大根判据和近似  $F$  统计量以及对每个因变量的单变量方差分析。除检验假设外，GLM 多因变量分析还产生参数估计。

通常使用 Priori 对比执行假设检验。当  $F$  检验已经表明显著性后，还可以使用 Post Hoc 检验评价指定均值间的差异。对边际均值的估计给出单元格预测均值的估计，这些

边际均值图很容易将某些关系可视化。对每个因变量分别进行 Post Hoc 多重比较检验。

残差、预测值、Cook 距离、杠杆值可以作为新变量保存在数据文件中，以便验证假设。可以求残差的 SSCP 矩阵，它是残差平方和和叉积的矩阵，残差协方差矩阵是残差的 SSCP 矩阵除以残差自由度。还有残差相关矩阵，这是标准化的残差协方差矩阵。

WLS Weight 选项允许指定一个变量，给观测测量不同的权重用于加权最小平方分析，或用做对不同测量精度的补偿。

为了检验有关参数估计的假设，GLM Multivariate 过程假设：

模型中的观测量和因变量之间的误差值是彼此独立的，一个好的研究设计一般要避免违反这个假设。

因变量的协方差在各单元中是常数。当单元（因素变量水平组合）尺寸（所包括的观测测量数）不同时，这一点尤其重要。

因变量的误差方差在因素变量水平的各组合中是相等的，即误差方差具有齐性。

### 9.4.2 多因变量方差分析过程和数据要求

#### 1. 多因变量方差分析的数据

多因变量方差分析的因变量应该是数值型尺度变量即连续变量。

因素变量是分类变量，可以是数值型，也可以是变量值小于或等于 8 个字符的字符型。分类预测因素可以选择作为模型中的自变量。因素的每个水平可以与因变量的值有不同的线性效应。GLM Multivariate 过程假设所有的模型因素都是固定的；也就是说，通常按照设计，它们被认为是所有感兴趣的变量值都出现在数据文件中。

协变量是与因变量相关的数值型变量。

因变量数据是多元正态分布的随机样本。在总体中，方差—协方差矩阵对所有单元都是相等的。要检验假设，可以用方差齐性检验（包括 Box's M 检验）和用 Spread-versus-Level 图，还可以检验残差和做残差图。

在进行方差分析之前，有必要使用 Explore 过程探索数据。对单个因变量，使用 GLM Univariate 进行方差分析。如果在不同情况下对每个被试对象测试同一个因变量，可以使用 GLM Repeated Measures 进行重复测量的方差分析。

#### 2. 操作方法

按 Analyze→General Linear Model→Multivariate 顺序单击菜单项，展开 Multivariate 主对话框，如图 9-22 所示。

多因变量方差分析过程与单因变量方差分析过程操作相同的有以下内容，方法见 9.3 节中 Model 功能（设计分析模型）、Contrasts 功能（选择对照方法）、Plots 功能（设定边际均值图参数）、Post Hoc 功能（选择多重比较方法）、Save 功能（选择要保存的输出变量）。虽然操作相同，输出结果却不同。因为有多个因变量，对每个因变量有一组输出。例如，在 Plot 对话框中指定了一个三维边际均值图形表达式，则对每个因变量，均按该



表达式生成一组边际均值图。

多因变量方差分析过程与单因变量方差分析过程操作上的不同之处在于，前者在 **Dependent Variable** 因变量框中可以选择多个因变量，而后者只能选择一个因变量。再有 **Options** 功能也与单因变量方差分析过程略有不同。

在主对话框中，单击 **Options** 按钮，展开如图 9-23 所示 **Multivariate: Options** 对话框。

(1) **Estimated Marginal Means** 栏中选项要估计的边界均值。操作方法见 9.3.2 节。

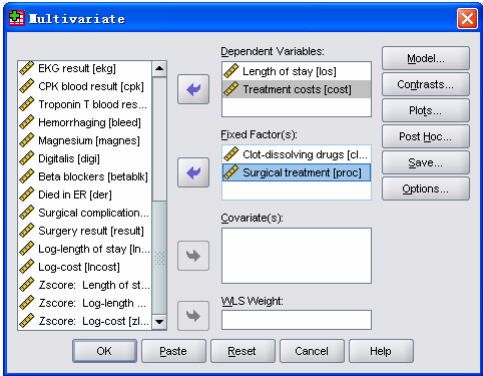


图 9-22 多因变量线性方差分析主对话框

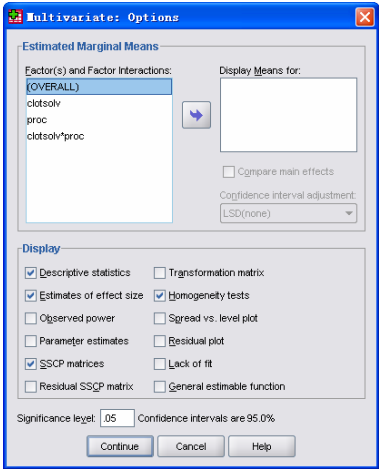


图 9-23 Options 对话框

(2) **Display** 框，选择输出项。

- **Descriptive statistics**，输出描述统计量，有观测量的均值、标准差和每个单元格中的观测量数。
- **Estimates of effect size**，输出效应量估计。选择此项，给出  $\eta^2$ （eta square）值。它是由一个自变量所解释的变异（SSH）对自变量解释的变异和未计入模型解释的变异总和（SSH+SSE）之比（SSH/（SSH+SSE））。
- **Observed power**，给出 F 检验的概率，它检验的是组间差异，在假设是基于观测值时，检验各种假设的功效。计算功效的显著性水平，系统默认临界值为 0.05。
- **Parameter estimates**，给出了各因素变量的模型参数估计、标准误、T 检验的 *t* 值、显著性概率和 95% 的置信区间。
- **SSCP matrices**，对每个效应给出平方和与叉积矩阵。对设计中的每个效应给出假设的和误差的 SSCP 矩阵。每个组间效应有不同的 SSCP 矩阵，对所有组间效应只有一个误差矩阵。
- **Residual SSCP matrix**，给出 RSSCP 残差的平方和与叉积矩阵。RSSCP 的维度与模型中因变量数相同。残差的协方差矩阵为 RSSCP 除以残差自由度。残差相关矩阵是由残差协方差矩阵标准化得来的。
- **Transformation matrix**，显示对因变量的转换系数矩阵或 M 矩阵。

• **Homogeneity tests**, 给出方差齐性检验结果。Levene 检验每个因变量在所有因素的水平组合间方差是否相等。

• **Spread vs. level plot**, 绘制观测量单元均值一标准差图和观测量单元均值一方差图。

• **Residuals plot**, 绘制残差图。给出观测值\*预测值\*标准化残差图。

• **Lack of fit**, 检查独立变量和非独立变量间的关系是否被充分描述。执行一种拟合不足检验（它要求对一个或几个自变量重复观测）如果检验被拒绝就意味着当前的模型不能充分说明响应变量与预测因素之间的关系，可能有变量被忽略，或是模型中需要其他项。

• **General estimable function**, 产生表明估计函数一般形式的表格。可以根据一般估计函数通过 **LMATRIX** 子命令自定义假设检验。

(3) 在 **Significance level** 框中, 改变 **Confidence intervals** 框内多重比较的显著性水平。

### 9.4.3 多因变量线性模型方差分析实例

【例 9】本例数据是对 1481 个心梗患者的数据，数据标号 data09-08。

1. 变量说明见表 9-40。作为对心梗（MI 或 心脏病发作）的初步治疗，有时在手术之前给溶解血栓（凝块溶解药）的药物，帮助清理患者的动脉。三种可用的药物是 **alteplase** 阿替普酶、**reteplase** 瑞替普酶和 **streptokinase** 链激酶。阿替普酶和瑞替普酶是新药，较昂贵。一个地区的卫生保健系统想确定，是否他们的价格一效应足够代替链激酶。溶解血栓的药物有一个好处就是外科手术比较平稳，因而痊愈周期比较短。如果新药是有效的，患者住院的时间就会较短。地区的卫生保健系统希望，较短的住院时间将有助于补偿较大的术前新药的花费。

表 9-40 变量说明

变量名	标 签	中文标签	值	值标签	值标签(中文)
los	Length of stay	住院时间长短			
cost	Treatment costs	治疗花费			
clotsolv	Clot-dissolving drugs	凝块溶解药	1	Streptokinase	链激酶
			2	reteplase	瑞替普酶
			3	alteplase	阿替普酶
proc	Surgical treatment	手术治疗	1	PTCA	经皮冠状动脉成型术
			2	CABG	搭桥术

数据文件中包括接受溶解血栓药物治疗的心梗患者 1481 个的样本的处理记录。使用 **GLM Multivariate** 对住院时间（天）和治疗药物的花费进行多元方差分析。

#### 2. 初步分析操作步骤

(1) 读取数据 data09-08。按 **Analyze**→**General Liner Model**→**Multivariate** 顺序单击菜单项，打开 **Multivariate Contrasts** 主对话框。

(2) 在主对话框中定义分析变量:

① 将住院时间变量 `los` 和治疗花费变量 `cost` 移入 **Dependent Variable** 框, 作为因变量。

② 将 `clotsolv` 和 `proc`(凝块消溶药和手术治疗)变量作为固定因素移入 **Fixed Factor(s)** 栏内。

(3) 由于要使用系统默认的全模型, 因此 **Model** 对话框不用打开。

(4) 单击 **Contrasts** 按钮, 打开对话框如图 9-24 所示。

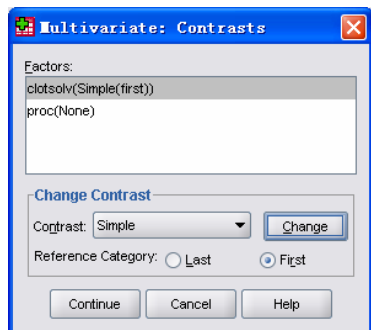


图 9-24 设置对比参数对话框

① 选择 `clotsolv(None)` 作为对比变量。

② 在 **Change Contrast** 栏内, 单击 **Contrast** 参数框内向下箭头, 展开比较方法表, 选择 **Simple** 作为比较类型, 再选择 **First** 项为比较参考类。然后单击 **Change** 按钮。在 **Factor** 栏内显示的对比表达式为: `clotsolv(simple(first))`。

(5) 在 **Options** 对话框中选择: **Descriptive statistics** 描述统计量; **Estimates of effect size** 估计效应大小; **SSCP matrices** 显示 **SSCP** 矩阵; **Homogeneity tests** 齐性检验;

(6) 单击 **Continue** 按钮, 返回主对话框。单击

**OK** 按钮, 提交系统执行。

3. 与以上选择操作对应的命令语句如下。

#### GLM

```
los cost BY proc clotsolv                                ①
/CONTRAST (clotsolv)=Simple(1)                            ②
/METHOD = SSTYPE(3)                                    ③
/INTERCEPT = INCLUDE                                   ④
/PRINT = DESCRIPTIVE ETASQ TEST(SSCP) HOMOGENEITY        ⑤
/CRITERIA = ALPHA(.05)                                   ⑥
/DESIGN = proc clotsolv proc*clotsolv .                  ⑦
```

① GLM语句调用GLM过程。

② 指定因变量为 `lose`、`cost`, **BY**后面是固定因素变量`proc`和`clotsolv`。

③ 对因素变量`clotsolv` 分组的因变量均值进行**Simple**比较, 即都与第一水平比较。

④ 模型包括截矩。

⑤ 输出项包括描述统计量、效应估计、**SSCP**矩阵和齐性检验结果。

⑥ 置信水平为 $\alpha=0.05$ 。

⑦ 模型为两个主效应`proc`、`clotsolv`一个交互效应`proc*clotsolv`。

4. 输出结果：见表 9-41 至表 9-43。

表 9-41 为组间因素各水平组合的单元频数。可以看出单元大小不一。

表 9-42 为每个因变量按服用的凝块消溶药和手术类型分组的描述统计量。

表 9-43 是多元检验的 SSCP 矩阵。多元检验表对每个模型效应显示了四种显著性检验。

• Pillai's Trace 是一个正值统计量。该统计量值越大表明对模型贡献的效应越多。

• Wilks' Lambda 是一个正值统计量，其值在 0~1 之间。统计量值越小表明效应对模型贡献越多。

• Hotelling's Trace 是检验矩阵特征值之和。它是一个正值统计量。值越大表明对模型贡献的效应越多。Hotelling's Trace 永远大于 Pillai's Trace，但当检验矩阵的特征值很小时，这两个统计量接近相等。这表明效应对模型没什么贡献。

• Roy's Largest Root 是检验矩阵的最大特征值，它是一个正值统计量，其值越大表明贡献给模型的效应越多。Roy's Largest Root 永远小于或等于 Hotelling's Trace。当这两个统计量相等时该效应主要与一个因变量相联系，在因变量之间存在很强的相关性，或者该效应对模型没有什么贡献。

可以看出凝块消溶药物对模型贡献不大，因为他们的 Pillai's trace、Hotelling's trace、Roy's Largest Root 的值分别为 0.026、0.027、0.027 都很小，而 Wilks' Lambda 的值却很大，0.974 接近 1。

以上 4 个多元统计量都转换到具有近似或确切的 F 分布的检验统计量，给出了 F 分布的假设自由度（分子）、误差自由度（分母）以及显著性概率。主效应 clotsolv 和 proc 的显著性值 sig 小于 0.05，表明效应对模型的贡献。

相比之下，它们的交互项对模型没有贡献。虽然 clotsolv 对模型有贡献，因为 Pillai's Trace 的值与 Hotelling's Trace 的值接近，他的贡献不是很大。更直接的方法是看 Partial Eta Squared 偏  $\eta^2$  统计量。该统计量报告每一项的实际的显著性，是根据由效应计算的变异

表 9-41 各单元频数

Between-Subjects Factors		
	Value Label	N
Surgical treatment	1 PTCA	907
	2 CABG	574
Clot-dissolving drugs	1 Streptokinase	116
	2 Reteplase	696
	3 Alteplase	669

表 9-42 描述统计量

Descriptive Statistics					
	Surgical treatment	Clot-dissolving drugs	Mean	Std. Deviation	N
Length of stay	PTCA	Streptokinase	4.94	1.105	68
		Reteplase	4.81	1.072	441
		Alteplase	4.68	1.048	398
		Total	4.77	1.066	907
	CABG	Streptokinase	7.25	1.263	48
		Reteplase	6.62	1.137	255
		Alteplase	6.48	1.135	271
		Total	6.60	1.163	574
	Total	Streptokinase	5.90	1.633	116
		Reteplase	5.47	1.399	696
		Alteplase	5.41	1.396	669
		Total	5.48	1.422	1481
Treatment costs	PTCA	Streptokinase	28.3838	3.27388	68
		Reteplase	29.6674	3.18096	441
		Alteplase	29.8073	3.60094	398
		Total	29.6326	3.39406	907
	CABG	Streptokinase	44.7225	5.42780	48
		Reteplase	44.6251	5.22506	255
		Alteplase	44.7432	5.63081	271
		Total	44.6890	5.42789	574
	Total	Streptokinase	35.1447	9.14344	116
		Reteplase	35.1476	8.27021	696
		Alteplase	35.8575	8.62337	669
		Total	35.4681	8.50314	1481

与由效应和剩在误差里的效应之和的比值。偏  $\eta^2$  的值较大的表明较大的模型效应，最大值为 1。由于 clotsolv 的偏  $\eta^2$  非常小，无论是对住院时间长短还是对治疗花费，变量 clotsolv 的偏  $\eta^2$  都非常小，分别为 0.015 和 0.02，说明它对模型的贡献不是很大。

表 9-43 多元检验的 SSQP 矩阵

Multivariate Tests <sup>a</sup>						
Effect		Value	F	Hypothesis df	Error df	Partial Eta Squared
Intercept	Pillai's Trace	.975	28781.280 <sup>a</sup>	2.000	1474.000	.975
	Wilks' Lambda	.025	28781.280 <sup>a</sup>	2.000	1474.000	.975
	Hotelling's Trace	39.052	28781.280 <sup>a</sup>	2.000	1474.000	.975
	Roy's Largest Root	39.052	28781.280 <sup>a</sup>	2.000	1474.000	.975
proc	Pillai's Trace	.622	1212.157 <sup>a</sup>	2.000	1474.000	.622
	Wilks' Lambda	.378	1212.157 <sup>a</sup>	2.000	1474.000	.622
	Hotelling's Trace	1.645	1212.157 <sup>a</sup>	2.000	1474.000	.622
	Roy's Largest Root	1.645	1212.157 <sup>a</sup>	2.000	1474.000	.622
clotsolv	Pillai's Trace	.026	9.833	4.000	2950.000	.013
	Wilks' Lambda	.974	9.892 <sup>a</sup>	4.000	2948.000	.013
	Hotelling's Trace	.027	9.952	4.000	2946.000	.013
	Roy's Largest Root	.027	19.909 <sup>b</sup>	2.000	1475.000	.026
proc * clotsolv	Pillai's Trace	.004	1.508	4.000	2950.000	.002
	Wilks' Lambda	.996	1.508 <sup>a</sup>	4.000	2948.000	.002
	Hotelling's Trace	.004	1.509	4.000	2946.000	.002
	Roy's Largest Root	.004	3.022 <sup>b</sup>	2.000	1475.000	.004

a. Exact statistic  
b. The statistic is an upper bound on F that yields a lower bound on the significance level.  
c. Design: Intercept+proc+clotsolv+proc \* clotsolv

相比较而言，proc 的偏  $\eta^2$  非常大，分别为 0.291 和 0.621，这正是所期望的。外科手术是患者必须接受的治疗，导致在医院住的时间的效应和最终花费比服用溶解血栓剂更大。多元检验也表明 clotsolv 效应是显著的 sig=0.000 即小于 0.01，这意味着，至少有一种药的效应与其他不同。对比的结果将表明是不同的。

表 9-44 均值比较的结果

表 9-44 是均值比较的结果，显示了三种凝块消溶药分组的住院时间长短的均值比较和治疗花费的均值比较。变量 clotsolv 药物的第一水平 streptokinase 链激酶是指定的参考类，可以看出第二组与第一组组均值比较结果说明使用 streptokinase 链激酶比使用 reteplase 瑞替普酶多住院 0.382 天，少花费 593 美金。由于对逗留医院时间长度的显著性值为 0.001，小于 0.05，因此可以推断这个差异不是偶然的。costs 治疗花费的显著性值大于 0.10，所以，这个差异完全可能由于随机变异引起。

Contrast Results (K Matrix)				Dependent Variable	
Clot-dissolving drugs				Length of stay	Treatment costs
Simple Contrast <sup>a</sup>					
Level 2 vs. Level 1	Contrast Estimate	Hypothesized Value	Difference (Estimate - Hypothesized)	-.382	.593
				0	0
				-.382	.593
	95% Confidence Interval for Difference	Lower Bound	Upper Bound	.112	.439
				.001	.176
				-.602	-.267
				-.162	1.453
Level 3 vs. Level 1	Contrast Estimate	Hypothesized Value	Difference (Estimate - Hypothesized)	-.516	.722
				0	0
				-.516	.722
	95% Confidence Interval for Difference	Lower Bound	Upper Bound	.112	.439
				.000	.100
				-.736	-.138
				-.296	1.583

a. Reference category = 1

第二个对比比较了第三个水平与第一水平, 即 alteplase 阿替普酶的效应和 streptokinase 链激酶的效应。平均上, 患者服用阿替普酶比服用链激酶大约少住院 0.516 天, 治疗费用高出 722 美金。由于 Length of stay 在医院逗留时间的显著性值小于 0.05, 可以推断该差异并非偶然。Treatment costs 的显著性值大于 0.10, 所以该差异可能完全是由于随机变异引起的。

综上所述, 使用 alteplase 和 reteplase 似乎减少患者住院时间。此外, 该项减少足以弥补治疗的费用。这样, alteplase 和 reteplase 的使用应该排在 streptokinase 的前面。然而在采用这个计划之前, 应该证明模型的假设检验是准确无误的。

表 9-45(a)是协方差矩阵的齐性检验结果。检验的假设是因变量遵循多元正态分布, 而且, 方差-协方差矩阵在各种效应之间形成的单元是相等的。Box's M 检验的零假设是因变量协方差矩阵在各组之间是相等的。Box's M 检验统计量被转换为具有  $df_1$  和  $df_2$  自由度的  $F$  统计量。这里的检验的  $p$  值小于 0.05, 拒绝原假设。这样模型的结果是不可信的。

Box's M 对大数据集是敏感的, 意味着当有大量观测数据时, 即使偏离齐性很小也可以检测出来。此外, 对偏离正态假设也很敏感。

表 9-45(b)是 Levene's 检验。其假设是各因素水平组合所定义的单元之间误差变异相等。对每个因变量分别进行检验。Length of stay 住院时间的  $p$  值大于 0.10, 因此不足以在这个检验中拒绝零假设(不排除在更多样本时, 或另一个检验方法时拒绝零假设)。Treatment costs 治疗花费的  $p$  值小于 0.05, 表明对这个变量, 违反了变异相等的假设。像 Box's M 一样, Levene's 检验对大数据集是敏感的。由于齐性检验的结果违反了进行多元方差分析的假设, 无法得出可信的结论。为什么? 怎样才能得出可信的结论呢?

表 9-45 协方差矩阵的齐性检验

Box's Test of Equality of Covariance Matrices

Box's M	270.509
F	17.908
df1	15
df2	358296.5
Sig.	.000

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups

a. Design: Intercept+proc+clotsolv+proc \* clotsolv

(a)

Levene's Test of Equality of Error Variances

	F	df1	df2	Sig.
Length of stay	1.507	5	1475	.185
Treatment costs	10.001	5	1475	.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+proc+clotsolv+proc \* clotsolv

(b)

## 5. 检查分析条件

(1) 为解决问题, 作出直方图, 粗略查看一下因变量的正态性。

作图步骤是按 Graphs→Legacy Dialog→histogram 顺序单击菜单项, 打开做直方图的对话框。将变量 los 住院时间长短移到 Variable 栏内。选择 display normal curve 要求显

示正态曲线作为比较参考。在输出窗口显示的直方图如表 9-25(a)所示。使用同样方法作因变量 **cost** 治疗花费的直方图如图 9-25 (b)所示。图 9-27 住院时间和治疗花费的直方图。可以看出住院时间变量近似正态分布，而治疗花费有明显的两个峰，每个都近似正态分布。一元方差分析表明，这是由于不同手术类型之间花费差异显著造成的。

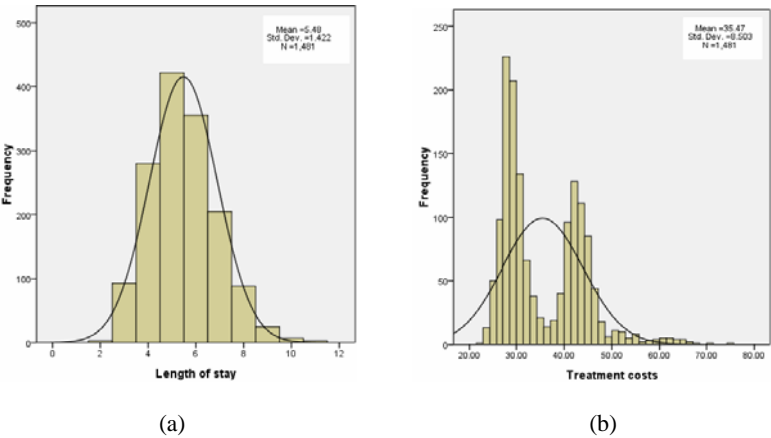


图 9-25 住院时间和手术花费直方图

(2) 进一步分别粗略检查分类变量 **proc**，外科手术两个分类 **CABG** 搭桥术和 **PTCA** 冠状动脉成型术的花费是否是正态分布。

操作步骤是使用 **Split File** 功能将数据按 **proc** 的分类分开，按 **Data→Split File** 顺序打开分割文件对话框，如图 9-26 所示。选择 **Compare groups**，并将变量 **proc(surgical treatment)**移到 **Groups Based on** 栏内，单击 **OK** 按钮。然后按上述步骤作直方图。两类手术的花费直方图如图 9-27 所示。

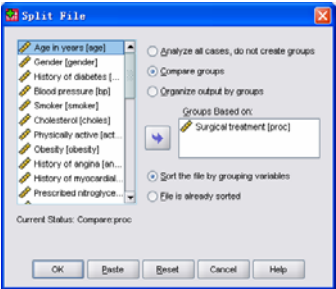


图 9-26 分割数据文件对话框

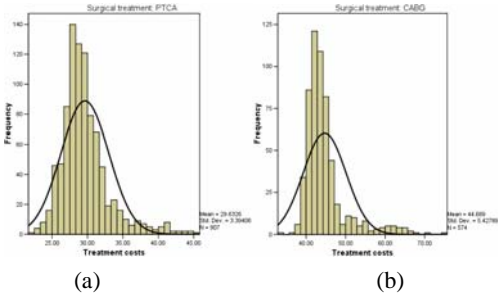


图 9-27 PTCA 和 CABG 的手术花费分布图

(3) 虽然两种手术类型的花费都是近似正态分布的，但是从图中可以看出都左偏，因此为进一步改善正态性，对治疗花费变量 **cost** 进行常用对数转换。使用 **Transform** 菜单中的 **Compute** 命令，利用 **LG10** 函数生成新变量 **logcost=LG10(cost)**。

(4) 进行两因变量单因素的方差分析。步骤如下：

① 仍然让数据文件处于按手术类型 `proc` 分开状态，以便并列比较。

② 按 `Analyze→General Linear Model→Multivariate` 顺序单击菜单，打开主对话框，将因变量 `los` 住院天数、`logcost` 治疗花费的对数变量作为因变量移到 `Dependent Variable` 栏中。将 `clotsolv` 凝块溶解药作为因素变量送入 `Fixed Factor(s)` 栏内。

③ 在 `Contrasts` 对话框中，选择以变量 `clotsolv` 的第一水平作为参考类的简单比较。

④ 在 `Options` 对话框中将 `OVERALL`、`clotsolv` 送入 `Display Means for` 栏中，并选择 `Compare main effect`。在 `Display` 栏中选择以下输出项：`Descriptive statistics`，要求显示描述统计量；`Estimates of effect size`，要求估计的效应；选中 `Homogeneity test`，要求进行方差齐性检验。

⑤ 在 `Post Hoc` 多重比较对话框中选择对 `clotsolv` 变量进行 `Post Hoc` 检验。

在 `Equal Variances Assumed` 方差相等假设栏中选择 `Tukey`，在 `Equal Variances Not Assumed` 栏中选择 `Dunnett's T3`。

(5) 执行的程序如下：

<code>SORT CASES BY proc .</code>	①
<code>SPLIT FILE</code>	②
<code>LAYERED BY proc .</code>	
<code>GLM</code>	③
<code>los logcost BY clotsolv</code>	
<code>/CONTRAST (clotsolv)=Simple(1)</code>	
<code>/METHOD = SSTYPE(3)</code>	
<code>/INTERCEPT = INCLUDE</code>	
<code>/POSTHOC = clotsolv ( TUKEY T3 )</code>	
<code>/EMMEANS = TABLES(OVERALL)</code>	
<code>/EMMEANS = TABLES(clotsolv) COMPARE ADJ(LSD)</code>	
<code>/PRINT = DESCRIPTIVE ETASQ HOMOGENEITY</code>	
<code>/CRITERIA = ALPHA(.05)</code>	
<code>/DESIGN = clotsolv .</code>	

程序解释：

① 是变量 `proc` 手术类型排序。

② 是按 `proc` 变量将数据文件分割为两页，以便后面的程序均对变量 `proc` 的各水平分开进行分析。在后面的程序中，`proc` 变量不应该出现在分类变量的位置上，例如不能作为因素变量参与分析。

③ `GLM` 过程中的第一语句设置了因变量是 `los` 和 `log cost`，因素变量是 `clotsolv`。二元单因素的方差分析。



(6) 主要输出结果在表 9-46~表 9-56 中。

表 9-47 描述统计量

Descriptive Statistics

Surgical treatment		Clot-dissolving drug	Mean	Std. Deviation	N
PTCA	Length of stay	Streptokinase	4.94	1.105	68
		Reteplase	4.81	1.072	441
		Alteplase	4.68	1.048	398
		Total	4.77	1.066	907
	logcost	Streptokinase	1.4504	.04802	68
		Reteplase	1.4700	.04392	441
		Alteplase	1.4715	.04883	398
		Total	1.4692	.04671	907
CABG	Length of stay	Streptokinase	7.25	1.263	48
		Reteplase	6.62	1.137	255
		Alteplase	6.48	1.135	271
		Total	6.60	1.163	574
	logcost	Streptokinase	1.6478	.04796	48
		Reteplase	1.6470	.04538	255
		Alteplase	1.6478	.04910	271
		Total	1.6474	.04731	574

表 9-46 组间因素各单元频数

Between-Subjects Factors

Surgical treatment		Clot-dissolving drug	Value Label	N
PTCA	Clot-dissolving drugs	1	Streptokinase	68
		2	Reteplase	441
		3	Alteplase	398
CABG	Clot-dissolving drugs	1	Streptokinase	48
		2	Reteplase	255
		3	Alteplase	271

(7) 分析与结论

因为数据文件按手术类型分开了，产生的输出表格都是按手术类型并列的，单独得出结论。这与手术类型变量作为一个因素的结果是不同的。

表 9-46 组间因素各单元频数。可以看出各单元中观测量数是不相等的。

表 9-48 协方差矩阵相等的 Box 检验

Box's Test of Equality of Covariance Matrices

PTCA	Box's M	12.208
	F	2.021
	df1	6
	df2	246014.589
CABG	Sig.	.059
	Box's M	7.804
	F	1.288
	df1	6
	df2	126223.460
	Sig.	.259

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.  
a. Design: Intercept+clotsolv

表 9-49 误差方差相等的 Levene 检验

Levene's Test of Equality of Error Variances

Surgical treatment		F	df1	df2	Sig.
PTCA	Length of stay	.545	2	904	.580
	logcost	1.950	2	904	.143
CABG	Length of stay	.524	2	571	.592
	logcost	.820	2	571	.441

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

表 9-47 按凝块溶解药分组的因变量描述统计量。作为后面分析的参考数据。

表 9-48 是 Box 检验的结果，检验的零假设是：因变量的协方差矩阵在 clotsolv 不同的凝块溶解的各组中相等。无论是 PTCA 经皮冠状动脉成型术还是 CABG 搭桥术的 Sig 值是计算的 F 值大于  $\alpha = 0.05$  的临界值的概率  $p$ ，表中的该值均大于 0.05，证据不足以在这个检验中拒绝零假设。

表 9-50 对效应的四种检验

Multivariate Test Results							
Surgical treatmen		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
PTCA	Pillai's trace	.038	8.824	4.000	1808.000	.000	.019
	Wilks' lambda	.962	8.889 <sup>a</sup>	4.000	1806.000	.000	.019
	Hotelling's trace	.040	8.955	4.000	1804.000	.000	.019
	Roy's largest root	.038	17.317 <sup>b</sup>	2.000	904.000	.000	.037
CABG	Pillai's trace	.038	5.484	4.000	1142.000	.000	.019
	Wilks' lambda	.962	5.528 <sup>a</sup>	4.000	1140.000	.000	.019
	Hotelling's trace	.039	5.571	4.000	1138.000	.000	.019
	Roy's largest root	.039	11.165 <sup>b</sup>	2.000	571.000	.000	.038

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

以上两个检验结果对因素水平组合形成的单元中观测测量数不相等的情况非常重要。

表 9-49 是 Levene 检验的结果。检验的零假设是因变量在 clotsolv 不同的凝块溶解药的各组中的误差方差相等。是对两个因变量分别进行的检验。从表中的 Sig 值可以看出两种手术的花费对数和住院天数的检验结果都不足以在这个检验中拒绝零假设。

表 9-51 多元方差分析检验结果

Tests of Between-Subjects Effects								
Surgical treatme	Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
PTCA	Corrected Model	Length of stay	5.725 <sup>a</sup>	2	2.863	2.529	.080	.006
		logcost	.026 <sup>b</sup>	2	.013	6.112	.002	.013
	Intercept	Length of stay	10695.320	1	10695.320	9448.846	.000	.913
		logcost	989.845	1	989.845	458854.4	.000	.998
	clotsolv	Length of stay	5.725	2	2.863	2.529	.080	.006
		logcost	.026	2	.013	6.112	.002	.013
	Error	Length of stay	1023.254	904	1.132			
		logcost	1.950	904	.002			
	Total	Length of stay	21624.000	907				
		logcost	1959.676	907				
	Corrected Total	Length of stay	1028.979	906				
		logcost	1.976	906				
CABG	Corrected Model	Length of stay	24.504 <sup>c</sup>	2	12.252	9.316	.000	.032
		logcost	7.47E-005 <sup>d</sup>	2	3.73E-005	.017	.984	.000
	Intercept	Length of stay	14546.869	1	14546.869	1061.280	.000	.951
		logcost	858.818	1	858.818	382466.4	.000	.999
	clotsolv	Length of stay	24.504	2	12.252	9.316	.000	.032
		logcost	7.47E-005	2	3.73E-005	.017	.984	.000
	Error	Length of stay	750.931	571	1.315			
		logcost	1.282	571	.002			
	Total	Length of stay	25800.000	574				
		logcost	1559.156	574				
	Corrected Total	Length of stay	775.436	573				
		logcost	1.282	573				

a. R Squared = .006 (Adjusted R Squared = -.003)

b. R Squared = .013 (Adjusted R Squared = -.011)

c. R Squared = .032 (Adjusted R Squared = .028)

d. R Squared = .000 (Adjusted R Squared = -.003)

表 9-52 不同凝块消溶药对住院时间和治疗花费的影响（PTCA）

Surgical treatment=PTCA			
Clot-dissolving drugs Simple Contrast <sup>a</sup>		Dependent Variable	
		Length of stay	logcost
Level 2 vs. Level 1	Contrast Estimate	-.129	.020
	Hypothesized Value	0	0
	Difference (Estimate - Hypothesized)	-.129	.020
	Std. Error	.139	.006
	Sig.	.351	.001
	95% Confidence Interval for Difference	Lower Bound Upper Bound	-.401 .008
			.143 .031
Level 3 vs. Level 1	Contrast Estimate	-.258	.021
	Hypothesized Value	0	0
	Difference (Estimate - Hypothesized)	-.258	.021
	Std. Error	.140	.006
	Sig.	.065	.001
	95% Confidence Interval for Difference	Lower Bound Upper Bound	-.532 .009
			.016 .033

a. Reference category = 1

表 9-53 不同凝块消溶药对住院时间和治疗花费的影响（CABG）

外科手术 = 搭桥术

凝块消溶药 Simple Contrast <sup>a</sup>		Dependent Variable	
		住院时间(天)	治疗花费 对数(底10)
Level 2 vs. Level 1	Contrast Estimate	-.634	-.001
	Hypothesized Value	0	0
	Difference (Estimate - Hypothesized)	-.634	-.001
	Std. Error	.180	.007
	Sig.	.000	.922
	95% Confidence Interval for Difference	Lower Bound Upper Bound	-.989 .015
			-.280 .014
Level 3 vs. Level 1	Contrast Estimate	-.774	.000
	Hypothesized Value	0	0
	Difference (Estimate - Hypothesized)	-.774	.000
	Std. Error	.180	.007
	Sig.	.000	.999
	95% Confidence Interval for Difference	Lower Bound Upper Bound	-1.127 .015
			-.421 .015

a. Reference category = 1

表 9-50 是多元检验的结果，四种检验的统计量，Pillaise Trace 、Hotelling's Trace、Roy's Largest Root 统计量越大对模型贡献越大，但是表中值都很小，Wilks lambda 统计量的值越小对模型贡献越大，而表中相应的值却很大，接近 1。所以四个统计量都说明因素变量的 clotsolv 效应对模型的贡献不大。但是表中的 F 检验的 Sig 值，即大于 F 临界值的概率都小于 0.01，而偏  $\eta^2$ （Partial Eta Square）值也都很小，又说明他们是有贡献的，但贡献不大。

表 9-51 是多元方差分析表。对住院时间（天）变量检验的零假设是：不同的凝块消溶药组的平均住院时间（天数）之间无显著差异。

对治疗花费对数变量检验的假设是：不同的凝块消溶药组的平均治疗花费（以 10 为底的对数）之间无显著差异。

表 9-54 均值多重比较表

Pairwise Comparisons								
Surgical treatment	Dependent Variable	(I) Clot-dissolving drugs	(J) Clot-dissolving drugs	Mean Difference (I-J)	Std. Error	Sig. <sup>a</sup>	95% Confidence Interval for Difference <sup>a</sup>	
							Lower Bound	Upper Bound
PTCA	Length of stay	Streptokinase	Reteplase	.129	.139	.351	-.143	.401
			Alteplase	.258	.140	.065	-.016	.532
		Reteplase	Streptokinase	-.129	.139	.351	-.401	.143
			Alteplase	.128	.074	.081	-.016	.273
		Alteplase	Streptokinase	-.258	.140	.065	-.532	.016
			Reteplase	-.128	.074	.081	-.273	.016
	logcost	Streptokinase	Reteplase	-.020*	.006	.001	-.031	-.008
			Alteplase	-.021*	.006	.001	-.033	-.009
		Reteplase	Streptokinase	.020*	.006	.001	.008	.031
			Alteplase	-.001	.003	.646	-.008	.005
		Alteplase	Streptokinase	.021*	.006	.001	.009	.033
			Reteplase	.001	.003	.646	-.005	.008
CABG	Length of stay	Streptokinase	Reteplase	.634*	.180	.000	.280	.989
			Alteplase	.774*	.180	.000	.421	1.127
		Reteplase	Streptokinase	-.634*	.180	.000	-.989	-.280
			Alteplase	.140	.100	.163	-.057	.336
		Alteplase	Streptokinase	-.774*	.180	.000	-1.127	-.421
			Reteplase	-.140	.100	.163	-.336	.057
	logcost	Streptokinase	Reteplase	.001	.007	.922	-.014	.015
			Alteplase	6.20E-006	.007	.999	-.015	.015
		Reteplase	Streptokinase	-.001	.007	.922	-.015	.014
			Alteplase	-.001	.004	.861	-.009	.007
		Alteplase	Streptokinase	-6.20E-006	.007	.999	-.015	.015
			Reteplase	.001	.004	.861	-.007	.009

Based on estimated marginal means

\*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 9-55 住院天数的一致性子集

Surgical treatment=PTCA			
	Clot-dissolving drugs	N	Subset
			1
Tukey HSD <sup>a,b</sup>	Alteplase	398	4.68
	Reteplase	441	4.81
	Streptokinase	68	4.94
	Sig.		.085

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = 1.132.

a. Uses Harmonic Mean Sample Size = 153.957.

b. Alpha = .05.

Surgical treatment=CABG				
	Clot-dissolving drugs	N	Subset	
			1	2
Tukey HSD <sup>a,b</sup>	Alteplase	271	6.48	
	Reteplase	255	6.62	
	Streptokinase	48		7.25
	Sig.		.650	1.000

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = 1.315.

a. Uses Harmonic Mean Sample Size = 105.467.

b. Alpha = .05.

先看外科手术是冠状动脉成型术（PTCA）这一组，分类变量 clotsolv 的 F 检验的显著性概率，对住院时间（天）Sig=0.08，大于 0.05，不足以拒绝原假设。说明不同的凝块消溶药组的平均住院时间之间无显著性差异。对治疗花费对数 Sig=0.02 小于 0.05，拒绝原假设。说明在本例条件下，不同的凝块消溶药组的治疗花费均值差异显著。

再看外科手术是搭桥术（CABG）这一组，分类变量 clotsolv 的 F 检验的显著性概率，对住院时间（天）Sig=0.00 小于 0.05，拒绝原假设说明不同的凝块消溶药组的平均住院时间之间有显著性差异。表中对治疗花费对数 Sig=0.982，即  $p>0.05$ ，不足以拒绝原假设。说明在本例条件下，没有证据显示，不同的凝块消溶药组的治疗花费均值没有显著差异。

表9-56 治疗花费（对数）的一致性子集

Surgical treatment=PTCA

		N	Subset	
			1	2
Tukey HSD <sup>a,b</sup>	Clot-dissolving drugs			
	Streptokinase	68	1.4504	
	Reteplase	441		1.4700
	Alteplase	398		1.4715
	Sig.		1.000	.958

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = .002.

a. Uses Harmonic Mean Sample Size = 153.957.

b. Alpha = .05.

Surgical treatment=CABG

		N	Subset	
			1	
Tukey HSD <sup>a,b</sup>	Clot-dissolving drugs			
	Reteplase	255	1.6470	
	Alteplase	271	1.6478	
	Streptokinase	48	1.6478	
	Sig.			.993

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = .002.

a. Uses Harmonic Mean Sample Size = 105.467.

b. Alpha = .05.

表 9-52 是做经皮冠状动脉成型术一组不同凝块消溶药对住院时间和治疗花费影响的对比表。

第二水平与第一水平比较即使用新药瑞替普酶与使用链激酶相比，平均住院时间少了 0.129 天，Sig=0.351 说明这个差异是由随机因素引起的，具有一定的偶然性。平均治疗花费的对数高出 0.020，表中 Sig=0.001，即  $p$  值小于 0.05，说明不是偶然的。转换后为高出  $10^{0.02}=1.047$  千美金。

第三水平与第一水平比较即使用阿替普酶与使用链激酶相比，平均住院时间少了 0.258 天，Sig=0.065 说明这个差异是由随机因素引起的，具有一定的偶然性。平均治疗花费的对数高出 0.021，输出表中 Sig=0.001 即  $p$  值小于 0.05，说明不是偶然的。转换后为高出  $10^{0.021}=1.05$  千美金。

表 9-53 是做搭桥术一组不同凝块消溶药对住院时间和治疗花费影响的对比表。

第二水平与第一水平比较即使用新药瑞替普酶与使用链激酶相比，平均住院时间少了 0.634 天，Sig 小于 0.01 说明这个差异不是随机因素引起的。根据描述统计量表中的总平均住院天数是 6.6 天，使用新药使住院时间减少了近 10%。平均治疗花费的对数高出 0.001，Sig=0.922，即  $p$  大于 0.05，说明是偶然的。两种药的平均治疗花费是相等的。

第三水平与第一水平比较即使用阿替普酶与使用链激酶相比，平均住院时间少了 0.774 天，Sig 小于 0.01 说明这个差异不是由随机因素引起的。使用新药阿替普酶使住院时间减少了 11.3%。平均治疗花费的对数高出 0，表中 Sig=0.999。即  $p>0.05$ ，没有证据拒绝两种药的平均治疗花费是相等的假设。

表 9-54 是均值多重比较结果，带有“\*”标记的是差异显著的两个水平的均值。

表 9-55 是住院天数的一致性子集,对经皮冠状动脉成型术(PTCA)来说,见左表,无论使用哪种凝块消溶药的住院天数均属于同一子集;对搭桥术(CAGB)来说,见右表,两种新药的住院天数属于同一子集,链激酶属于单一子集,平均住院天数高于两组使用新药的。

表 9-56 是治疗花费的一致性子集,对经皮冠状动脉成型术(PTCA)来说两种新药的治疗花费属于同一子集,链激酶属于单一子集,平均治疗花费低于两组使用新药的。对搭桥术来说,无论使用哪种凝块消溶药的治疗花费均属于同一子集。

结论:该研究项目提出的对于 MI 患者服用新的凝块消溶药,比使用链激酶是否可以减少住院天数弥补治疗的高昂费用呢?

搭桥术,服用新药可以缩短住院时间 10%~11%,没有证据说明治疗花费的差异常。

成型术,服用新药的住院时间与服用链激酶是一致的,花费要高出一千多美金。

结论是对于心梗患者做搭桥手术的可以使用新的凝块消溶药 Alteplase 阿替普酶或 Reteplase 瑞替普酶代替原来常用的链激酶。可以缩短住院时间而不增加治疗费用。对做经皮冠状动脉成型术的患者使用新药只能增加治疗费用,不能缩短住院时间。

以上结论也可以从多重均值比较表 9-54 和表 9-55、表 9-56 两对一致性子集表得出。

## 9.5 重复测量设计的方差分析

### 9.5.1 重复测量方差分析概述

#### 1. 重复测量方差分析的概念与重复测量方差分析的过程

最简单的重复测量方差分析是对实验对象的两次测量,例如实验前后各测量一次,分析实验前后样本均值间差异的显著性,从而推断实验所施加的处理或不同条件的效应。使用配对样本 T 检验进行分析。也可以进行相关分析。本章介绍的是当测量次数大于等于 3 的情况下的方差分析方法。

GLM 重复测量属于高级分析过程,是对同一因变量进行重复测量,可以是同一条件下进行的重复测量,目的在于研究各种处理之间是否存在显著性差异的同时,研究被试者之间的差异;也可以是不同条件下的重复测量,目的在于研究各种处理间是否存在显著性差异的同时,研究形成重复测量条件间的差异以及这些条件与处理间的交互效应。

例如在对某种动物不同种系的繁殖实验中,使用两种种系的动物每种系若干只,在不同温度下,测试其体重、胎儿重、脂肪厚度等。可以分析种系之间(组间因素)有关繁殖指标的差异,研究不同温度(组内因素)下繁殖指标间的差异以及随温度变化各指标的变化趋势。

GLM Repeated Measures 重复测量过程是对每个观测对象在不同条件下进行几次相同的测量的方差分析。使用这个一般线性模型过程,可以检验组间因素(处理)的效应

和组内（重复测量）因素的效应的零假设，可以检验处理因素的效应以及重复测量因素间的交互效应。另外，还包括协变量效应，也包括组间因素的与协变量之间的交互效应。

## 2. 几个术语

**Between-Subjects Factor** 组间因素，即处理因素。组间因素的水平把观测量划分成几个组。这里的组间因素的水平是指处理的不同水平。重复测量过程研究不同水平间因变量之间差异。

**Within-Subjects Factor** 组内因素。组内因素形成重复测量条件。组内因素的不同水平决定了对观测对象的重复测量次数。重复测量过程研究重复测量的各组间的差异。

**Measure** 测量。即模型中的因变量，是对每次测量的量给一个名字。

**Covariates** 协变量。尺度类型的预测因子如果在因素水平的组合（单元）中，与因变量的值是线性相关的，应该选做模型中的协变量。

**Interactions** 交互效应。**GLM Repeated Measures** 过程默认产生具有全部因素交互效应的模型，这意味着因素水平的每个组合与因变量有不同的线性效应。另外，如果认为在协变量与因变量之间的线性关系因因素水平的不同而不同时，可以指定因素变量与协变量的交互效应。

例如，在一项减肥研究中，每周测量几个人的体重，共测量 5 周。在数据文件中，每个人是一个观测对象，或称事件。各周所测量的体重记录在变量 `weight1` ~ `weight5` 中。每个人的性别记录在另一个变量中。体重对每个观测对象重复测量，可以通过定义组内因素组织起来。该因素可叫做 `week`。定义它有 5 个水平。在主对话框中，变量 `weight1`, ..., `weight5` 被分派成 `week` 的 5 个水平。数据文件中，性别变量可以定义为组间变量，以便研究男女之间的差异。`weight` 作为 **Measures** 变量，该测量在数据文件中并不作为变量存在，但是在这里定义。有时具有多个测量的模型叫做双重重复测量模型。

## 3. 偏差平方和的分解

在重复测量设计的方差分析中的偏差平方和分解如下。以  $m$  水平的处理因素把样本观测量分为  $m$  组， $j=1\sim m$ ；每组有  $n$  个实验对象， $i=1\sim n$ ；对每个实验对象进行 1 次测量， $k=1\sim l$  的重复测量实验为例。

(1) 总处理的偏差平方和被分解为处理间的偏差平方和、重复测量间的偏差平方和与处理因素与重复测量因素之间的交互的偏差平方和。

$$S_{\text{总处理}} = n \times \sum_{j=1}^m \sum_{k=1}^l (\bar{x}_{jk} - \bar{\bar{x}})^2 \quad \text{自由度: } m \times l - 1$$

式中

$$\bar{x}_{jk} = \frac{1}{n} \sum_{i=1}^n x_{ijk} \quad \bar{\bar{x}} = \frac{1}{m \times n \times l} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^l x_{ijk}$$

$$S_{\text{重复测量间}} = n \times m \sum_{k=1}^l (\bar{x}_k - \bar{\bar{x}})^2 \quad \text{自由度为 } l - 1,$$

$$\text{其中, } \bar{x}_k = \frac{1}{l} \sum_{i=1}^n \sum_{j=1}^m x_{ijk}$$

$$S_{\text{处理组间}} = n \times l \sum_{j=1}^m (\bar{x}_j - \bar{x})^2 \quad \text{自由度为 } m-1$$

$$\text{其中, } \bar{x}_j = \frac{1}{n \times l} \sum_{i=1}^n \sum_{k=1}^l x_{ijk}$$

$$S_{\text{处理因素*重复测量因素}} = S_{\text{总处理}} - S_{\text{处理组间}} - S_{\text{重复测量组间}}$$

交互项的自由度为  $(m-1)*(l-1)$

(2) 整个样本的总的偏差平方和分解为体现个体间变异的, 各次重复测量合计的偏差平方和与体现重复测量间差异的重复测量组内偏差平方和。公式如下

$$S_{\text{总}} = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^l (x_{ijk} - \bar{x})^2 \quad \text{自由度为 } n*m*l-1$$

$$S_{\text{组间合计}} = \frac{1}{k} \sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \bar{x}_{ij})^2 \quad \text{自由度为 } m*n-1$$

$$\text{式中, } \bar{x}_{ij} = \frac{1}{n \times m} \sum_{i=1}^n \sum_{j=1}^m x_{ij}; \text{ 重复测量的合计: } x_{ij} = \sum_{k=1}^l x_{ijk};$$

$$S_{\text{组内合计 (重复测量间)}} = S_{\text{总处理}} - S_{\text{组间合计 (观测对象间)}} \quad \text{自由度为 } m*(n-1)$$

(3) 上述的组间合计的偏差平方和分解为处理组间的偏差平方和与组间误差的偏差平方和。计算公式如下:

$$S_{\text{组间误差}} = S_{\text{组间合计}} - S_{\text{处理组间}} \quad \text{公式见上面。自由度为 } m*(n-1)$$

(4) 上述的组内合计的偏差平方和分解为重复测量间的偏差平方和、交互作用的偏差平方和与组内误差的偏差平方和。前三项根据前面公式可以计算出, 故

$$S_{\text{组内误差}} = S_{\text{组间合计}} - S_{\text{处理因素*重复测量因素}} \quad \text{公式见上面。自由度为 } m*(n-1)*(l-1)$$

(4)、(5)中的偏差平方和除以各自的自由度得到相应的均方。它们与误差均方之商即为F检验的F值。

#### 4. SPSS 中重复测量方差分析的假设检验

假设有  $k$  个样本, 即是对同一组观测对象在  $k$  个条件下的重复测量。

(1) 原假设  $H_0$ :  $k$  次重复测量的样本均数都相同即  $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k = \mu$ ,  $k$  个样本有共同的方差  $\sigma$ 。则  $k$  个样本来自具有共同的方差  $\sigma$  和相同的均数  $\mu$  的总体。SPSS 将  $k$  次重复测量样本看作  $k$  个因变量做四种多元检验。如果经过检验  $F$  值远远大于临界值,  $p < 0.05$ , 推翻原假设,  $p > 0.05$  无法拒绝原假设, 样本来自相同总体, 处理无作用。即  $k$  次重复测量之间无显著差异。

(2) 如果定义了组间因素变量, 在重复测量方差分析中, 组间偏差平方和即是反映了



该分组变量各水平间的差异。检验的零假设是  $H_0$ ：该分类变量各水平组成的样本来自均值相同的总体。如果经过计算，组间均方远远大于误差均方， $F$  值远远大于临界值， $F_b > F_{0.05df_b, df_{rse}}$ ，则  $p < 0.05$ ，推翻原假设，说明分组变量各水平的因变量均值差异显著，样本来自不同的正态总体，否则无法拒绝原假设。不足以在这个检验中拒绝零假设（不排除在更多样本时，或另一个检验方法时拒绝零假设）。

5. 趋势分析

如果重复测量的条件是有序变化的，例如是在不同时间点或不同温度点下进行的测量，可以分析因变量均值随时间变化的趋势或随温度变化的趋势是线性的、二次或者为三次的或更高次的。

9.5.2 重复测量方差分析的数据文件结构

1. 重复测量设计的数据及数据文件结构

在实验中进行重复测量的因变量应该是等间隔测度的（连续的）数值型数据。这些重复测量的因变量可以是在不同条件下的，对同一组观测对象进行的测量结果，组合后作为组内因素，这是重复测量设计必需的。

在实验中的分类变量体现观测对象的分组，不同组观测对象不同，在方差分析中作为组间因素。最简单的方差分析中可以不包括组间因素。

要进行重复测量方差分析，数据的组织与其他类型的方差分析有所不同，它要求对被试者的若干次重复测试结果作为不同因变量出现在数据文件中。例如教育心理研究中的对刺激反应时测量的实验方法的研究中，设置了三个级别的视觉刺激作为处理因素变量，4 位被试者均接受三个级别的刺激，每个被试者给予一个编号，该变量不参与分析，只为输入数据及核对时使用。对每个被试者在同样条件下测试三次，原始实验数据记录见表 9-57，数据编号为 data09-09。

在数据窗口中建立数据文件的变量说明：number 被试者编号，vsno 视觉刺激等级（1=刺激 1、2=刺激 2、3=刺激 3），time1 反应时测量 1，time2 反应时测量 2，time3 反应时测量 3。数据文件中的数据样例见表 9-28，数据文件的结构对重复测量方差分析很重要，一定要把每次测量作为一个变量，否则无法使用 SPSS 的重复测量方差分析功能对数据进行分析。

表 9-57 重复测量的原始数据

受试者	刺激 1				刺激 2				刺激 3			
	1	2	3	4	5	6	7	8	9	10	11	12
反应时测量 1	0.9	1.5	0.5	0.8	2.4	1.9	2.9	2.4	1.5	2.1	1.1	1.6
反应时测量 2	1.2	1.1	0.8	1.3	2.8	2.4	3.3	2.8	1.2	1.9	1.5	1.8
反应时测量 3	0.7	0.8	0.5	0.9	2.1	2.2	2.7	2.9	1.9	2.2	1.0	1.3

	number	vsno	time1	time2	time3
1	1	1	.9	1.2	.7
2	2	1	1.5	1.1	.8
3	3	1	.5	.8	.5
4	4	1	.8	1.3	.9
5	1	2	2.4	2.8	2.1
6	2	2	1.9	2.4	2.2
7	3	2	2.9	3.3	2.7
8	4	2	2.4	2.8	2.9
9	1	3	1.5	1.2	1.9
10	2	3	2.1	1.9	2.2
11	3	3	1.1	1.5	1.0
12	4	3	1.6	1.8	1.3

图 9-28 重复测量数据文件结构

## 2. 重复测量方差分析的假设条件

重复测量设计中的每一元每组测量中的观测值应该是独立的，并符合多元正态分布，这与一元（ANOVA）、多元（MANOVA）分析一样。如果违反多元正态分布或观测值独立的假设，可能得到不可解释的结果。

重复测量设计还要求满足球形假设，即每组之间的方差-协方差矩阵相等，但在各观测值相等的情况，对该假设条件的要求并不严格。

### 9.5.3 组内因素的设置与重复测量方差分析过程

重复测量方差分析的功能模块调用步骤如图 9-2 所示。即按 Analyze→General Linear Model→Repeated Measures 顺序单击菜单项，打开 Repeated Measures Define Factor(s)对话框，如图 9-29 所示。

#### 1. 组内因素的定义

**注意，这里定义的不是数据文件中的变量。而是重复测量的变量组的代号。**

##### (1) 定义组内因素

当在菜单中选择了 Repeated Measures 时，并不马上展开重复测量方差分析的主对话框，而是需要先显示定义组内因素的对话框，如图 9-29 所示。

本例中研究的组内因素是由三个视觉刺激反应时构成的，这个组内因素命名为 time，共有三个水平。

① 在 Repeated Measures Define Factor(s)对话框内 Within-Subject Factor Name 框中输入组内因素名 time，代替原显示的 Factor1。

② 在 Number of Levels 框中输入因素水平数 3，如图 9-29 所示。输入结束后，Add 按钮加亮，单击该按钮，定义表达式显示在大矩形框中。如果研究的课题中还有另外的组内因素，可以用同样方法继续定义。

③ 已经定义的组内因素倘若有错误，单击出错的定义表达式，此时 Change、Remove 按钮加亮。单击 Change 按钮，表达式分解为组内因素名和水平，两部分分别显示在 Within-Subject Factor Name 和 Number of Levels 框中。在框中修改后，单击 Change 按钮，正确的表达式将显示在大矩形框中。删除已经定义并显示在大矩形框中的组内因素，可以单击该表达式，然后单击 Remove 按钮。

④ 如果对每个组内因素所代表的变量的测量仍有重复，在定义因素对话框 Define Factor(s)的下半部分，定义表示重复测试的变量。见图 9-29。

例如在减肥的研究中，20 个人为实验对象。在 5 周中，每周测一次体重。在数据文件中每个人是一个观测值（一个记录，占一行）。每周测得的体重数为 5 个变量 weight1~weight5 的值。另外用 gender 变量记录他（她）们的性别。对每个人来说，体重就是重复测量的。把 weight1~weight5 定义为 Within-Subject Factor Name，可以命名为 week，它有 5 个水平。在 Within-Subject Factor Name 栏中输入 week，在 Number of Levels 栏输

入 5, 单击 Add 按钮。在主对话框中, weight1~weight5 用于给 week 的 5 个水平赋值。性别变量可以定义为 **Between Subject Factor**, 以便研究男、女在减肥实验中的差异。如果课题要在每天测一次脉搏和呼吸, 则应该单击 **Measures** 定义这些测量结果。

## (2) 进入重复测量方差分析的主对话框

检查所有定义的组内因素表达式, 正确无误后, 单击 **Define** 按钮, 结束组内因素定义工作, 进入 **Repeated Measures** 主对话框, 如图 9-30 所示。

主对话框中有 4 个矩形框:

- 左面一个矩形框显示了在数据文件中输入的所有变量。
- 右面一个矩形框显示了在组内因素定义对话框中定义的所有因素水平与测量的组合, 标有 **Within-Subjects Variables**, 其后的括号中是已经定义的组内因素名。
- 下面一个矩形框标有 **Between-Subjects Factor(s)**, 组间因素框。
- 最下面的矩形框标有 **Covariates**, 协变量框。

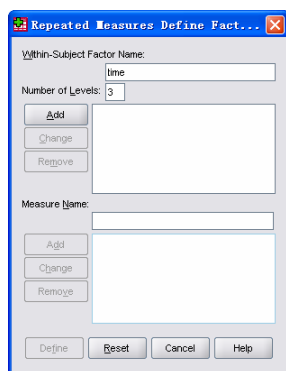


图 9-29 定义组内因素的对话框

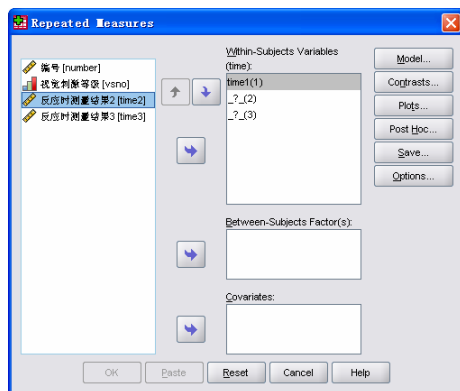


图 9-30 重复测量方差分析主对话框

## 2. 定义组内因素各水平组合与原始变量的对应关系

在组内变量矩形框中显示有一系列  $_{?}(n)$ , 表示组内变量第  $n$  个水平。

在原始变量表中选择读者认为是组内因素第  $n$  水平的变量并单击之。本例选择原始变量 **time1** 作为组内因素 **time** 的第一水平, 因此单击左面矩形框中的 **time1**, 然后单击向右箭头按钮。在右面矩形框中的 “ $_{?}(1)$ ” 变为 **time1(1)**。如果想要让 **time1** 作 **time** 变量的第二因素, 可以单击向下箭头按钮, 使 **time1(1)** 变为 **time1(2)**, 即可以使用上、下箭头按钮改变组内因素变量水平与原始变量的对应关系。

**注意:** 组内因素水平组合表达式的括号内是水平组合。本例只定义了一个组内因素 **time**, 因此表达式括号内为一个数字。如果定义了两个组内因素, 表达式括号内为两个用逗号隔开的水平序号。

例如, 如果定义了组内因素  $x$ , 有两水平;  $y$  有 3 水平。在组内因素矩形框中将出现

以下要求定义的表达式:

$\_?(1,1)$ 、 $\_?(1,2)$ 、 $\_?(1,3)$ 、 $\_?(2,1)$ 、 $\_?(2,2)$ 、 $\_?(2,3)$ 。

当然在数据文件中与之对应的应该有六个因变量。每个因变量对应着一种水平组合。读者不难反过来考虑这是怎样一种重复测量方差分析设计。

### 3. 定义方差分析的组间因素变量

在变量列表中选择组间因素变量,例如,选择 Vsno,送入 Between-Subjects Factor(s) 下面的矩形框中。

### 4. 定义协变量及其类型

如果有协变量,在变量框选中协变量,单击向右箭头按钮,送入 Covariates 框中。

### 5. 定义分析模型

在主对话框中,单击 Model 按钮,展开 Repeated Measures: Model 模型定义对话框,如图 9-31 所示。

(1) 在 Specify Model 栏中选择定义模型的方式

① Full Factorial, 饱和模型, 系统默认方式。

② Custom, 自定义方式。选择自定义方式激活 4 个框。可以定义组内模型。

#### (2) 自定义模型

在 Within-Subjects 矩形框中列出了组内因素变量。Between-Subjects 矩形框中列出了组间因素变量。对应的右面两个矩形框分别是 Within-Subjects Model 被试内模型和 Between-Subjects Model 被试间模型。中间的 Build Term(s) 栏是对应效应类型的下拉列表,其中有主效应、各级交互效应。选择一种类型使用,参见 9.3.2 小节中叙述的方法定义被试内模型和被试间模型。注意如果有两个以上协变量,不能指定协变量与协变量之间的交互效应。但可以使用 Transform 菜单中的 Compute 功能使两个协变量相乘建立新变量,建立指定新变量的各种效应。

#### (3) 选择计算组间模型平方和方法

在 Sum of Squares 框内可以选择分解平方和的方法。

6. 有关 Contrast 功能选择、Plots 功能选择、Post Hoc 功能选择、Save 功能选择、Options 功能选择,均与单因变量多因素方差分析的选项、含义相同。

只在 Options 对话框中的 SSCP matrices 项与单因变量多因素方差分析不同。

SSCP matrices 复选项对设计中每个效应给出平方和与叉积矩阵。单因变量多因素方

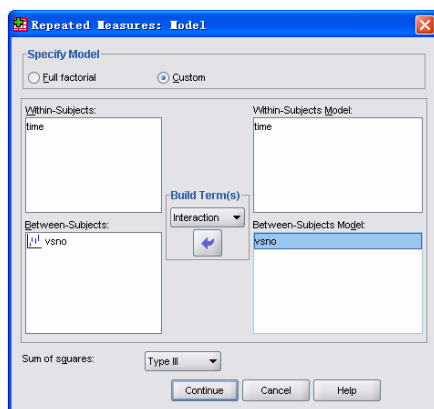


图 9-31 重复测量模型定义对话框

差分析中，对所有组间效应只给出一个误差阵。只有对重复测量，对每个组间效应既给出假设的 **SSCP** 矩阵，也给出误差 **SSCP** 矩阵。

7. 在各子对话框中的选项选定后，在各子对话框中，单击 **Continue** 按钮，返回主对话框。一切选择确定后，可以按下节实例操作。

9.5.4 重复测量方差分析实例

【例 10】下面以一元重复测量分析为例说明重复测量设计方差分析原理。研究四种药物对某生化指标的作用，5 名被试者参与实验，见表 9-58。每种药物实验之间相隔时间足以避免药物相互之间影响。这是无处理（条件）分组，一个组内因素的实验设计。数据编号为 data09-10。

研究的零假设为，四种药物对某生化指标作用（组内）无显著性差异。

表 9-58 不同药物对受试者的影响

受试者	药物 1	药物 2	药物 3	药物 4
1	30	28	16	34
2	14	18	10	22
3	24	20	18	30
4	38	34	20	44
5	26	28	14	30

1. 操作步骤

(1) 按 **Analyze→General Linear Model→Repeated Measures** 顺序单击菜单项，打开 **Repeated Measures Define Factor(s)**对话框。

(2) 定义组内因素

① 在 **Repeated Measures Define Factor(s)**

对话框的 **Within-Subjects Factors Name** 框中删除原有的 **Factor1**，输入组内因素名 **med**。注意这里的 **med** 不是数据文件中的变量名，是变量 **med1**、**med2**、**med3**、**med4** 这组重复测量的变量的代号。

② 在 **Number of Levels** 框中输入组内因素 **med** 的水平数“4”。

③ 单击 **Add** 按钮，在大矩形框中显示 **med(4)**，**Define** 按钮变亮，组内因素设置完成。由于没有再次的重复实验，不必操作 **Measures** 按钮。

④ 单击 **Define** 按钮，确认以上一个组内因素变量的单元。显示 **Repeated Measures** 主对话框。

(3) 在主对话框中设置分析变量

设置组内因素 **med** 与原始变量之间的对应关系：在左面矩形框中选择 **med1~med4**，4 个变量，单击向右箭头按钮，在 **Within-Subjects Factor(s)**栏中，框中第一项变为 **med1(1)**，第二个组内因素 **med2(2)**和第三、四个组内因素 **med3(3)**和 **med4(4)**。

(4) 本研究的重复测量实验设计没有考虑协变量，故无须对 **Covariates** 框进行操作。

(5) 根据数据特点，只能检验 4 种药物对生化指标变化的差异的显著性，以及受试者间的差异性，无须指定分析模型。根据实验目的无须进行均值比较。**Contrasts** 对话框的操作也无须进行。

(6) 输出选项

在主对话框中，单击 **Options** 按钮，展开选项对话框，为在输出中显示观测均值，

在 Factor(s) and Factor Interarctions 框中选择 med 项, 单击向右箭头按钮, 使 med 显示在 Display Means for 框中。选择 Descriptive Statistics 复选项。目的是要求输出各种药物作用下生化指标的描述统计量, 以便根据输出得出结论。

(7) 由以上各种选择确定的命令程序。单击“运行”按钮, 提交运行, 或在主对话框中单击 OK 按钮, 提交系统执行。

## 2. 命令程序与程序解释

(1) 在主对话框中, 单击 Paste 按钮, 在 Syntax 窗口中显示如下命令程序清单:

```
GLM                                     ①
    med1 med2 med3 med4
    /WSFACTORS =med 4 Polynomial        ②
    /METHOD =SSTYPE(3)                 ③
    /EMMEANS = TABLES(med)
    /PRINT = DESCRIPTIVE                ④
    /CRITERIA=ALPHA(.05)                ⑤
    /WSDSIGN= med.                      ⑥
```

(2) 命令程序解释:

① GLM 命令, 调用 GLM 过程。因变量 med1、med2、med3、med4;  
② 定义组内因子命令, 组内因素名为 med, 有 4 个水平。对应的因变量值为 med1 ~ med4。

③ 模型分析的分解平方和的方法为 Type III。

④ 要求输出描述统计量。

⑤ 显著性水平  $\alpha = 0.05$ 。

⑥ 重复测量方差分析的组内因素为 med。

3. 输出结果: 见表 9-59 至表 9-61。表 9-62 是为比较做的单因变量方差分析结果

## 4. 结果解释与分析

表 9-59(a)是组内因素基本数据信息。组内因素 med, 4 个水平, 作为 4 个因变量 med1, med2, med3 和 med4。

表 9-59(b)是重复测量变量的描述统计量。是每种药物作用后生化指标均值、标准差及观测量数。

表 9-59 基本信息与描述统计量

Within-Subjects Factors		Descriptive Statistics		
Measure: MEASURE_1				
med	Dependent Variable	Mean	Std. Deviation	N
1	med1	26.40	8.764	5
2	med2	25.60	6.542	5
3	med3	15.60	3.847	5
4	med4	32.00	8.000	5

(a)

(b)

表 9-60 为多变量检验结果。四种方法的 F 检验的概率 Sig 值均为 0.034 小于 0.05，说明四种药物对该生化指标的作用差异显著。常用的 Wilks' Lambda 应该是 0~1 之间的值，其值越接近 0，越拒绝作为因变量的四种药物对某生化指标作用无差异的假设。现在的值是 0.023，也说明了拒绝原假设的结论。Roy's Largest Root 是检验矩阵的最大特征值，其值越大表明贡献给模型的效应越多，本例中的值为 42.618。Roy's Largest Root 永远小于或等于 Hotelling's Trace。当这两个统计量相等时说明在因变量之间存在很强的相关性，以重复测量的 4 个变量 med1~med4 做相关分析，可以证明这个结论。Hotelling's Trace 永远大于 Pillai's Trace。这两个统计量相距越大，越表明该效应对模型贡献越多。而本例中的 Pillai's Trace 值为 0.977，Hotelling's Trace 值为 42.618。也说明因变量的四种药物对某生化指标这个因变量的贡献是比较多的。

表 9-60 药物间多元检验

Multivariate Test <sup>a</sup>					
Effect		Value	F	Hypothesis df	Sig.
med	Pillai's Trace	.977	28.412 <sup>a</sup>	3.000	.034
	Wilks' Lambda	.023	28.412 <sup>a</sup>	3.000	.034
	Hotelling's Trace	42.618	28.412 <sup>a</sup>	3.000	.034
	Roy's Largest Root	42.618	28.412 <sup>a</sup>	3.000	.034
a. Exact statistic					
b.					
Design: Intercept					
Within Subjects Design: med					

表 9-61 被试者内效应方差分析结果

Tests of Within-Subjects Effects						
Measure: MEASURE_1						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
med	Sphericity Assumed	698.200	3	232.733	24.759	.000
	Greenhouse-Geisser	698.200	1.815	384.763	24.759	.001
	Huynh-Feldt	698.200	3.000	232.733	24.759	.000
	Lower-bound	698.200	1.000	698.200	24.759	.008
Error(med)	Sphericity Assumed	112.800	12	9.400		
	Greenhouse-Geisser	112.800	7.258	15.540		
	Huynh-Feldt	112.800	12.000	9.400		
	Lower-bound	112.800	4.000	28.200		

表 9-62 一元完全随机设计方差分析

ANOVA					
服药物1后生化指标					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	698.200	3	232.733	4.692	.016
Within Groups	793.600	16	49.600		
Total	1491.800	19			

表 9-61 为对组内效应的方差分析结果即重复测量间的差异。第一行是在满足球形假设条件下，对 F 分子、分母自由度不做调整的条件下的检验结果。下面三行是在不满足球形假设时三种不同的检验方法，对 F 检验的分子分母自由度做了不同的调整的检验结果。四种条件下的 F 检验对应的 Sig 值，即  $p<0.05$ ，因此拒绝组内因素无差异的原假设。

说明被试者对不同药物的反应差异有统计意义上的显著。由于平衡设计对球形假设条件没有严格要求,而且在表中几种情况的检验结果都相同,故没有给出球形假设的输出表。

表 9-62 为按完全随机设计进行方差分析(根据数据文件 data09-10a)的结果。与按重复测量方差分析结果比较,可以看出,按重复测量方差分析的  $F$  值远远大于按完全随机设计来分析的  $F$  值,按重复测量方差分析显著性概率更远离 0.01。

可以看出,较少的实验对象反复使用,不但可以减少人力、财力在实验中的消耗,而且可以很好地减少由于实验对象个体偏差引起的误差方差。当然,需要避免的是两次实验间的相互影响。例如,本例中,对同一个实验对象给 4 种药物进行实验,两种药物实验间的时间间隔可能要相当长,以避免前一种药物对后一次实验的影响。这是专业问题,不是统计方法问题。需要读者注意。

### 9.5.5 关于趋势分析

#### 1. 趋势分析的概念

当重复测量的条件是某些顺序变量时,可以分析重复测量的因变量随顺序变量变化的趋势。

【例 11】选择 16 名实验对象,使用两种方法锻炼他们的记忆。训练一段时间后,每间隔一天测试一次记忆情况,共测试 5 次。每次测试对每个参与实验的人员均按一定法则打分。数据见 data09-11。这是一个组内因素、一个组间因素的重复测量设计的例题。因为组内因素是与时间有关的变量,因此不但可以分析比较两种训练记忆的方法哪个更有效,还可以得到随时间的推移,记忆分数随时间下降的数学模型。如果回忆的下降在整个测量的时间段上是个常数,则会发现记忆的下降与时间之间是线性关系。如果回忆的下降表现在前两天,第 3 天开始则急剧下降,会得到一个二次趋势。如果在第一天表现为下降,然后在以后的几天急剧下降,最后达到稳定,则回忆与时间的关系呈现为 3 次关系。见图 9-32。

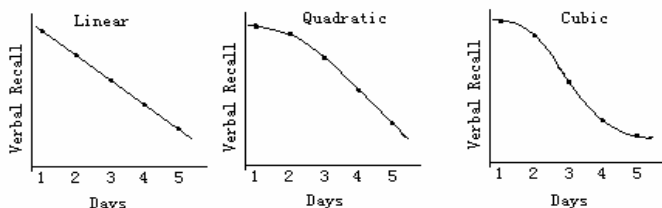


图 9-32 线性、二次、三次趋势图

#### 2. 有关记忆趋势分析的操作步骤

(1) 按 Analyze→General Linear Model→Repeated Measure 顺序单击菜单项,打开重复测量对话框。

(2) Within-subject Factor Name: days 设置重复测量变量集名为 days。Number of



Levels: 5 重复测量变量集 (被试内因素) 水平数为 5。单击 Add→Define。

(3) 在 Repeated Measure 主对话框中, 左栏选择五个因变量 day1~day5, 单击向右箭头按钮, 右栏显示 day1(1), day2(2), ..., day5(5)。将变量 group 送入 Between-Subjects Factor(s) 栏中作为组间变量。

(4) 单击 Model 按钮, 在模型对话框只选择 Custom 自定义模型, Build Term(s) 栏的菜单中选择 Main effects, 将 Within-Subjects 栏中的 days 作为组内因素送入 Within-Subject Model 栏中, 将 Between-Subjects 栏中的 group 变量送入 Between-Subject Model 栏中。单击 Continue 按钮返回主对话框。

(5) Plots 对话框中, 选择 days 作横轴变量送入 Horizontal Axis 栏, 将 group 送入 Separate Lines 栏中作为分线变量。单击 Plots 的 Add 按钮确定要输出的图形。

(6) Options 对话框中, 选择 days、group、OVERALL 送入 Display Means for 栏。在 Display 复选项中的 Descriptive Statistics 和 Estimate of effect size。

因组间变量 group 只有两个水平, 不能进行均值的多重比较。

### 3. 程序与程序说明

(1) 选项选择完后, 在对话框中单击 Paste 按钮后得到如下程序语句:

GLM

day1 day2 day3 day4 day5 BY group	①
/WSFACTOR = days 5 Polynomial	②
/METHOD = SSTYPE(3)	③
/PLOT = PROFILE( days*group )	④
/EMMEANS = TABLES(group)	⑤
/EMMEANS = TABLES(days)	⑤
/EMMEANS = TABLES(OVERALL)	⑤
/PRINT = DESCRIPTIVE ETASQ	⑥
/CRITERIA = ALPHA(.05)	⑦
/WSDESIGN = days	⑧
/DESIGN = group.	⑨

#### (2) 语句说明

- ① 调用 SPSS 的 GLM 过程, 分组变量 group, 因变量 day1~day5。
- ② 组内因素为组合变量 days, 共 5 个水平。
- ③ 计算平方和的方法采用 Type III。
- ④ 制图, 前变量是横轴变量, 组合变量 days 的每个水平在横轴上占一个刻度。后面变量是分线变量。即变量 group 的每个水平做一个折线。
- ⑤ 三个关键字相同的语句, 要求输出变量 group、days 各水平均值的估计值和总估计值。

- ⑥ 要求输出描述统计量。
- ⑦ 检验临界值设定为 0.05。
- ⑧ 模型设计的组内因素是组合变量 days。
- ⑨ 模型设计的组间因素是分类变量 group。

4. 输出结果见表 9-63 至表 9-68、图 9-33。

5. 输出结果说明

表 9-63(a)显示了组内因素 days 由 5 个因变量组成。表 9-63 (b)显示了组间因素是按实验方法分为两组，每组 8 个实验对象。

表 9-64 描述统计量，每个因变量按实验组对照组分组显示均值标准差，观测量数  $N$ 。

表 9-63 组内因素和组间因素清单

Within-Subjects Factors						Between-Subjects Factors	
Measure: MEASURE_1							
	day						N
	1	2	3	4	5		
Dependent Variable	day1	day2	day3	day4	day5		
						实验方	1
						法分组	2
							8
							8

(a)

(b)

表 9-64 组合变量各水平及总描述统计量

Descriptive Statistics				
实验方法分组	Mean	Std. Deviation	N	
一天后分数				
1	34.25	6.228	8	
2	35.00	5.928	8	
Total	34.63	5.886	16	
两天后分数				
1	30.88	6.728	8	
2	31.63	5.097	8	
Total	31.25	5.779	16	
三天后分数				
1	24.50	4.986	8	
2	24.88	4.704	8	
Total	24.69	4.686	16	
四天后分数				
1	19.13	5.592	8	
2	20.25	3.882	8	
Total	19.69	4.686	16	
五天后分数				
1	16.88	5.890	8	
2	15.25	5.651	8	
Total	16.06	5.639	16	

表 9-65 针对 5 天作为 5 个因变量进行的多元检验。检验的假设是 5 天每天得分的均值相等，各天得分与教法之间无交互作用。可以看出四种方法的  $F$  检验  $p$  值（表中 Sig）均为 0.00 小于 0.001，因此 5 天之间得分均值间差异显著，这一点从 Wilk's Lambda 值为 0.059 接近 0 也可以得出同样结论。同时看出，5 天与教法之间的四种检验结果  $p$  值（表中 Sig）都大于 0.05，说明教法与测试延续时间之间无交互效应。

表 9-65 多元检验结果

Multivariate Tests <sup>a</sup>						
Effect		Value	F	Hypothesis df	Error df	Partial Eta Squared
day	Pillai's Trace	.941	43.509 <sup>a</sup>	4.000	11.000	.000
	Wilks' Lambda	.059	43.509 <sup>a</sup>	4.000	11.000	.000
	Hotelling's Trace	15.821	43.509 <sup>a</sup>	4.000	11.000	.000
	Roy's Largest Root	15.821	43.509 <sup>a</sup>	4.000	11.000	.000
day * group	Pillai's Trace	.364	1.573 <sup>a</sup>	4.000	11.000	.249
	Wilks' Lambda	.636	1.573 <sup>a</sup>	4.000	11.000	.249
	Hotelling's Trace	.572	1.573 <sup>a</sup>	4.000	11.000	.249
	Roy's Largest Root	.572	1.573 <sup>a</sup>	4.000	11.000	.249

a. Exact statistic  
b.  
Design: Intercept+group  
Within Subjects Design: day

表 9-66 组内因素效应检验结果

Tests of Within-Subjects Effects						
Measure: MEASURE_1						
Source		Type III Sum of Squares	df	Mean Square	F	Partial Eta Squared
day	Sphericity Assumed	3832.925	4	958.231	135.268	.000
	Greenhouse-Geisser	3832.925	1.870	2049.340	135.268	.000
	Huynh-Feldt	3832.925	2.302	1664.885	135.268	.000
	Lower-bound	3832.925	1.000	3832.925	135.268	.000
day * group	Sphericity Assumed	19.175	4	4.794	.677	.611
	Greenhouse-Geisser	19.175	1.870	10.252	.677	.507
	Huynh-Feldt	19.175	2.302	8.329	.677	.535
	Lower-bound	19.175	1.000	19.175	.677	.425
Error(day)	Sphericity Assumed	396.700	56	7.084		
	Greenhouse-Geisser	396.700	26.184	15.150		
	Huynh-Feldt	396.700	32.231	12.308		
	Lower-bound	396.700	14.000	28.336		

表 9-67 组内因素多项式对比检验结果（趋势分析）

Tests of Within-Subjects Contrasts						
Measure: MEASURE_1						
Source	day	Type III Sum of Squares	df	Mean Square	F	Partial Eta Squared
day	Linear	3792.756	1	3792.756	193.729	.000
	Quadratic	1.290	1	1.290	.253	.623
	Cubic	33.306	1	33.306	12.850	.003
	Order 4	5.572	1	5.572	5.197	.039
day * group	Linear	7.656	1	7.656	.391	.542
	Quadratic	5.469	1	5.469	1.074	.318
	Cubic	3.906	1	3.906	1.507	.240
	Order 4	2.144	1	2.144	1.999	.179
Error(day)	Linear	274.088	14	19.578		
	Quadratic	71.313	14	5.094		
	Cubic	36.288	14	2.592		
	Order 4	15.013	14	1.072		

表9-68 边际均值表

2. 实验方法分组					3. day				
Measure: MEASURE_1					Measure: MEASURE_1				
实验方法分组	Mean	Std. Error	95% Confidence Interval		day	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound				Lower Bound	Upper Bound
1	25.125	1.762	21.345	28.905	1	34.625	1.520	31.365	37.885
2	25.400	1.762	21.620	29.180	2	31.250	1.492	28.050	34.450
					3	24.688	1.212	22.089	27.286
					4	19.688	1.203	17.107	22.268
					5	16.063	1.443	12.968	19.157

表 9-66 组内因素效应检验, 无论是否符合球形假设的前提条件, 四种方法计算的组合变量 days 四类偏差平方和相等, 只是根据不同方法调整了自由度, F 检验的结果, 原假设成立的概率均小于 0.001。说明五天之间的分数均值差异显著。days 与 group 交互效应不显著。

表 9-67 利用对比进行的趋势分析结果。假设是①分数均值随天数变化的趋势不具有线性特性; ②不具有二次特性; ③不具有三次特性; ④不具有四次特性。各种回归分析的方差分析表明不足以在这个检验中拒绝零假设, 因为① $p < 0.001$ , ② $p = 0.623 > 0.05$ , ③ $p = 0.003 < 0.01$ , ④ $p = 0.039 > 0.01$ , 但  $p < 0.05$ 。因此可以认为得分随时间变化的趋势符合线性或三次函数的特征。作为结论犯错误的概率  $< 0.01$ 。比较 Sig 值的大小, 可以选择认为变化趋势为线性下降的。

表 9-68 为边际均值表。左面一个表是按实验方法分组的均值、标准误和 95% 置信区间。右面的表是按时间顺序分组的均值、标准误和 95% 置信区间。

图 9-33 是每天平均分数图。每种方法一条折线。可以看出实验方法 2 随时间的推移记忆的得分下降趋势近似直线, 实验方法 1 的折线近似于三次曲线。与表 9-67 中趋势分析的综合结果也是相符合的。

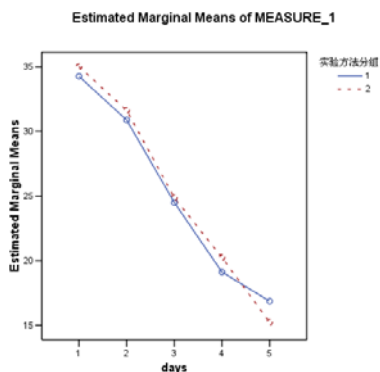


图 9-33 趋势图 (边际均值图)

## 9.6 方差成分分析

方差成分分析是研究混合效应模型中各随机效应对因变量变异的贡献。这个过程主要适用于对混合模型的分析, 如对裂区、单变量重复测量和随机区组设计的分析。通过计算方差成分, 可以找出减小方差的方向。方差成分分析过程共有 4 种分析方法: 最小正规二次无偏估计 (MINQUE)、方差分析 (ANOVA)、最大似然 (ML) 和有限最大似然法 (REML)。各种方法默认输出项都包括方差成分估计值。

如果使用最大似然和有限最大似然法还输出渐近协方差阵。其他输出还包括方差分析表和方差分析的期望均方，使用最大似然和有限最大似然法还输出迭代过程。

方差成分分析过程与 GLM 单因变量方差分析过程完全兼容。

**WLS Weight** 允许指定一个加权变量，进行加权分析时，用于给各观测量不同的权重。或作为不同测量精度的补偿。

方差成分分析要求因变量是数值变量。因素变量是分类变量，可以是数值型变量，也可以是最多由 8 个字符组成的字符型变量。至少要有一个因素是随机的。也就是说，该因素的水平必须是从可能的水平中随机采样得来的。协变量是数值型变量，并与因变量有一定的相关关系。

所有方差成分分析方法都假设，随机效应模型参数的均值为 0 和方差为有限常数，并且彼此不相关。不同效应的模型参数也不相关。

残差项也有零均值和有限常数方差。它与任意一个随机效应的模型参数都不相关。不同观测的残差项也假设为彼此不相关。

根据这些假设，随机因素同一水平的观测量是彼此相关的。ANOVA 和 MINQUE 方法不要求正态假设。虽然它们都是在正态假设条件下的方法，但它们都能缓解违反正态分布带来的影响。ML 和 REML 要求模型参数和残差项是正态分布的。

在进行方差成分分析之前，可以使用 Explore 过程检测数据。对假设检验，可以使用 GLM Univariate、GLM Multivariate 和 GLM Repeated Measures 几个过程进行。

### 9.6.1 方差成分分析过程

方差成分的功能模块调用步骤见图 9-2，即 Analyze→General Linear Model→Variance Components，最后展开 Variance Components 主对话框。如图 9-34 所示。

1. 定义因变量和随机因子。注意在做方差成分分析时一定要指定随机因子。在左边的

变量框中选因变量，单击向右箭头，将其送入 **Dependent Variable** 矩形框，然后再从左边的变量框选中随机因子，单击向右箭头，将其送入 **Random Factor(s)** 矩形框。

如果需要协变量分析，可以指定协变量进入 **Covariate(s)** 矩形框。如果需要分析权重，可以指定权重因子进入 **WLS Weight** 矩形框。完成以上工作后即可通过功能按钮展开相应的对话框，选择模型、选项，储存等内容。

2. 单击 **Model** 按钮，展开 **Model** 对话框，如图 9-35。选择分析模型。如果选择全模型，

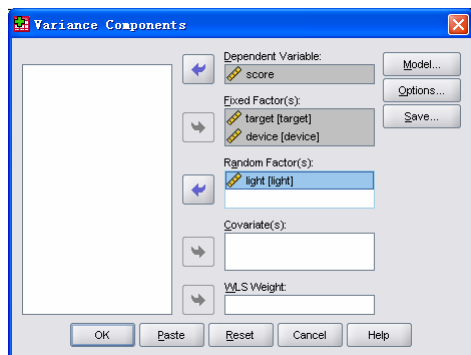


图 9-34 方差成分分析主对话框

模型中包括所有因素变量主效应、协变量主效应、因素变量之间的交互效应。不包括协变量交互项。如果选择自定义 Custom，可以指定因素变量与协变量的交互效应。模型中必须包括随机因素变量。

3. 单击 Options 按钮，展开 Options 对话框，选择分析方法，如图 9-36 所示。

(1) 在 Method 栏内指定一种进行方差成分分析的方法，有四种方法可供选择：

① MINQUE 选项，正态最小二次无偏估计，就固定效应而言，产生的估计是不变的。如果数据是正态分布的并且估计是正确的，使用此方法做方差大小估计，要比其他方法得到的方差小。这是系统默认方法。

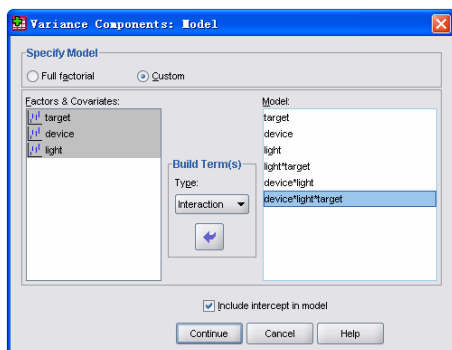


图 9-35 方差成分分析模型对话框

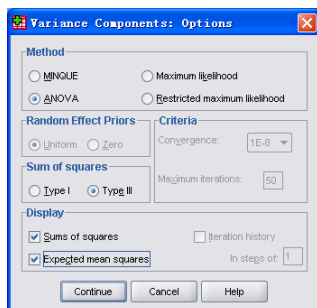


图 9-36 方差成分分析的选项对话框

② ANOVA (Analysis of Variance)选项，对每个效应使用 Type I 或 Type III平方和分解方法进行无偏估计。ANOVA 方法有时产生负方差估计，这表明模型不正确或估计方法不合适，或者需要更多的数据。

③ Maximum likelihood (ML)选项，最大似然法，使用迭代的方法产生与实际观测的数据最一致的估计，这些估计可能是有偏差的。该方法是接近正态的。ML 和 REML 估计在经转换后是不变的。该方法对固定效应作估计时未考虑自由度。

④ Restricted maximum likelihood (REML)选项，有限最大似然法。对许多平衡数据（并非对所有平衡数据），该方法比 ANOVA 法估计值要小。因为此方法对固定效应做调整，计算的标准误可能比 ML 法要小。在估计固定效应时考虑自由度。

(2) Random Effect Priors 栏，在系统默认 MINQUE 方法的同时，激活 Random Effect Priors 栏，也就是说该栏只对 MINQUE 方法有效。

① Uniform，指定此项意味着所有随机效应和残差项对观测量的影响相等。系统默认 Uniform 项。

② Zero，指定此项，假设随机效应方差相等且都为零。仅在指定了 MINQUE 方法时可以指定此选项。

(3) Sum of squares 栏, 在指定 ANOVA 方法的同时, 激活 Sum of squares 栏。

① Type I, 是系统默认的。用于分层模型方差成分的迭代。

② Type III, 仅用于 ANOVA 方法。

(4) Display 栏

① 在指定 ANOVA 方法的同时, 激活以下两个复选项。

- Sum of square, 要求显示平方和。

- Expected mean square, 要求显示期望均方值。

② 在指定 ML 或 REML 方法时, 只激活 Display 栏中的第三个选项: Iteration history 要求显示迭代过程。

(5) Criteria 栏指定 ML 或 REML 方法的同时才能激活该选择框。

① 在 Convergence 框中指定收敛判据, 下拉列表中以科学记数法列出选择范围 1E-6 到 1E-10, 表示  $10^{-6} \sim 10^{-10}$ , 共 5 个选项, 可以选择其中一个。

② 在 Maximum iterations 框中指定最大迭代次数。

4. 单击 Save 按钮, 展开 Save to New File 对话框, 如图 9-37 所示。该对话框可以将方差成分分析结果作为一个新的数据存储到指定的数据文件中, 以便进行其他的统计分析时使用。

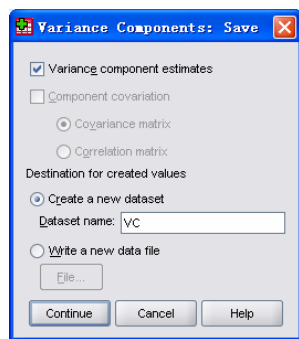


图 9-37 保存对话框

(1) 指定保存内容

① Variance component estimates, 指定保存方差成分估计值。

② Component covariation, 该项只有在选择 ML 或 REML 方法时才被激活。在以下两项中选择其一。

- Covariance matrix, 保存方差成分协方差矩阵。

- Correlation matrix, 保存方差成分相关矩阵。

(2) Destination for created values, 对所产生的方差成分估计值、和/或矩阵指定保存目标

① Create a new dataset, 可以保存成一个数据集, 但不是外部数据文件, 除非事先明确在这一个 SPSS 期间结束保存成数据文件。数据集只在当前的 SPSS 期间使用, 用作其他分析的数据。要在 Dataset name 后面给出数据集的名字。

② Write a new data file, 把分析结果产生的数据写入一个外部数据文件。

此项的选择激活 File 按钮, 单击该按钮, 展开 Variance Components: Save to File 对话框, 在该对话框中指定保存位置、文件类型和文件名。通常保存类型为 \*.sav 即 SPSS 数据文件类型。单击 Save 按钮, 返回主对话框。

### 9.6.2 方差成分分析实例

【例 12】本小节例题使用教育心理学实验中，心理运动测验分数与被试者必须瞄准的目标大小关系的资料，即 data09-07 数据。

#### 1. 操作步骤：

(1) 读取数据文件 data09-07。

(2) 按 Analyze→General Linear Model→Variance Components 顺序单击各菜单项，展开 Variance Components 主对话框，如图 9-34 所示。

(3) 定义因变量、因素变量和随机因素变量。在方差成分分析主对话框的变量列表中选择 score 作为因变量，单击向右箭头，将其送入 Dependent Variable 框内。因为被试者瞄准的目标和使用的设备是实验设计者选择的固定条件，所有研究者感兴趣的水平都包括在数据文件中了，属于固定因素。因此，在变量列表内选中 target、device 作为固定因素变量，单击第二个向右箭头按钮，将其送入 Fixed Factor(s)框中。由于认为亮度是随机地从亮度这个大总体中随机选择的两种亮度，可以认为亮度是随机因素，因此，选择 light 变量作为随机因素，单击第三个向右箭头按钮，将其送入 Random Factor(s)框中。

(4) 单击 Model 按钮，展开模型选择对话框，选择 Custom 自定义模型。

① 在 Build Term(s)框下确定 Main effects 主效应项，从左边的变量框分别选择 target、device、light 变量，单击向右箭头，送入 Model 框中。这样定义了三个主效应。

② 在 Builds Terms 框下确定 Interaction 交互效应项，从左边变量框选择 target 和 light 变量，单击向右箭头，同时送入 Model 框中，即确定 target\*light 交互项。用同样方法将 device\*light 和 target\*device\*light 送入 Model 框中。按 Continue 按钮，返回 Variance Components 主对话框。该步骤的目的是对主效应和与随机因素变量有关的交互效应做方差成分估计。

(5) 单击 Options 按钮，展开选择项对话框。

① 指定方差估计方法。在 Method 栏下，选中 ANOVA 法。

② 在 Sum of square 栏下选择 Type III 为分解偏差平方和的方法。

③ 在 Display 栏中选择显示 Sum of squares 平方和和 Expected mean square 期望均方。

单击 Continue 按钮返回主对话框。单击 Paste 按钮在语句窗口中显示生成的程序为：

VARCOMP	①
score BY light target device	②
/RANDOM = light	③
/METHOD = SSTYPE (3)	④
/PRINT = SS	⑤
/PRINT = EMS	⑥
/DESIGN =light target device device*light light*target device*light*target	⑦
/INTERCEPT = INCLUDE.	⑧



在语句窗口中，单击“运行”按钮，执行程序。

2. 语句解释

① 调用 VARCOMP 过程，进行方差成分分析。

② 主命令的一部分，指定因变量是 score，因素变量有三个：light、target、device。

③ RANDOM 子命令指定随机因素变量是 light。

④～⑥：METHOD 子命令指定使用 Type III 方法计算偏差平方和；PRINT 子命令指定要求输出平方和；PRINT 子命令指定要求输出期望均方。

⑦ DESIGN 子命令给出模型中指定要分析的效应：三个主效应，两个交互效应和一个三维交互效应。

⑧ INTERCEPT 子命令，要求模型中包括截距。

3. 运行结果见表 9-69 至表 9-75。

4. 运行结果解释：

表 9-69 为因素水平情况表，表中列出三个因素 target、device、light，每个因素的水平及值标签。表下方注明因变量是 score。

表 9-69 因素水平情况

Factor Level Information			
		Value Label	N
light	1	l1	60
	2	l2	60
target	1	t1	30
	2	t2	30
	3	t3	30
	4	t4	30
device	1	d1	40
	2	d2	40
	3	d3	40

Dependent Variable: score

表 9-70 为方差估计表。表中列出了各主效应和交互效应的平方和分解的结果，即各效应的偏差平方和值、各自的自由度，以及均方值。可以看出此表就是方差分析结果，不管将 light 变量作为固定因素还是随机因素，模型确定时，ANOVA 方差分析结果都是相同的，说明了方差成分分析与单因变量多因素分析是兼容的。这里的 Mean Square 也称作观测均方。

表 9-71 为方差成分表。列出各效应的方差估计值。注解中说明：

① 因变量为 score。

② 分析方法为 ANOVA，使用 Type III 方法计算偏差平方和。表中交互效应项 light\*device 的方差为负值，这是 ANOVA 方法可能发生的结果。其原因可能是：

- 指定的模型是错误的。
- 该方差估计的实际值为 0。

亮度与设备的交互效应可以不考虑，因为不但在方差成分表中显示了负值，在方差分析表中的均方值也是最小的，均方值仅有 6.3。见表 9-70 ANOVA 方差分析结果。

因此必须重新设计分析模型。进行第二次分析。只是在前面模型基础上，去掉 light\*device 交互项。结果见表 9-72 至表 9-74。

表 9-72 是第 2 次方差分析的 ANOVA 方差分析表。偏差平方和分解可以与表 9-70 比较，是去掉了 light\*devic 项，总偏差平方和值不变，该项的偏差平方和包括在 light\*device\*target 中了。

因变量 score 心理测试得分的方差，在交互项中，主要来源于随机变量 light 与固定因素变量 device、target 的三维交互效应。亮度与目标的交互效应不大。可以得出的结论是：亮度对测量得分的影响不能忽视，设备和目标是两个固定因素，是作为实验条件存在的。因此减小方差的方向要从减小它们与亮度的交互效应考虑。

表 9-70 ANOVA 方差分析结果

ANOVA			
Source	Type III Sum of Squares	df	Mean Square
Corrected Model	783.467	23	34.064
Intercept	3162.133	1	3162.133
target	235.200	3	78.400
device	86.467	2	43.233
light	76.800	1	76.800
light * target	93.867	3	31.289
light * device	12.600	2	6.300
light * target * device	278.533	12	23.211
Error	70.400	96	.733
Total	4016.000	120	
Corrected Total	853.867	119	

Dependent Variable: score

表 9-71 方差成分表

Variance Estimates	
Component	Estimate
Var(light)	1.040
Var(light * target)	.539
Var(light * device)	-.846 <sup>a</sup>
Var(light * target * device)	4.496
Var(Error)	.733

Dependent Variable: score

Method: ANOVA (Type III Sum of Squares)

- a. For the ANOVA and MINQUE methods, negative variance component estimates may occur. Some possible reasons for their occurrence are: (a) the specified model is not the correct model, or (b) the true value of the variance equals zero.

表 9-72 第二次方差分析

ANOVA			
Source	Type III Sum of Squares	df	Mean Square
Corrected Model	783.467	23	34.064
Intercept	3162.133	1	3162.133
target	235.200	3	78.400
device	86.467	2	43.233
light	76.800	1	76.800
light * target	93.867	3	31.289
light * target * device	291.133	14	20.795
Error	70.400	96	.733
Total	4016.000	120	
Corrected Total	853.867	119	

Dependent Variable: score

表 9-73 期望均方系数矩阵

Expected Mean Squares						
Source	Variance Component					Quadratic Term
	Var(light)	Var(light * target)	Var(light * device)	Var(light * target * device)	Var(Error)	
Intercept	60.000	15.000	20.000	5.000	1.000	Intercept, target, device, target * device
target	.000	15.000	.000	5.000	1.000	
device	.000	.000	20.000	5.000	1.000	
light	60.000	15.000	20.000	5.000	1.000	
light * target	.000	15.000	.000	5.000	1.000	
light * device	.000	.000	20.000	5.000	1.000	
light * target * device	.000	.000	.000	5.000	1.000	
Error	.000	.000	.000	.000	1.000	

Dependent Variable: score

Expected Mean Squares are based on Type III Sums of Squares.

For each source, the expected mean square equals the sum of the coefficients in the cells time variance components, plus a quadratic term involving effects in the Quadratic Term cell.

表 9-73 给出期望均方与方差成分之间的系数矩阵。方差成分分析的 ANOVA 方法是根据随机效应的期望均方与观测均方相等，来估计方差成分的。即根据表 9-72 和表 9-73 得出表 9-74 的结果。本例中：

$$EMS(\text{light} * \text{target} * \text{device}) = 5 * \text{Var}(\text{light} * \text{target} * \text{device}) + 1 * \text{Var}(\text{Error})$$

$$MS(\text{light} * \text{target} * \text{device}) = 20.795$$

$$\text{Var}(\text{Error}) = 0.733$$

$$\text{可以解出三阶交互效应的方差成分 } \text{Var}(\text{light} * \text{target} * \text{device}) = 4.012$$

$$EMS(\text{light} * \text{target}) = 5 * \text{Var}(\text{light} * \text{target} * \text{device}) + 15 * \text{Var}(\text{light} * \text{target}) + \text{Var}(\text{Error})$$

$$\text{可以解出随机效应的二阶交互效应的方差成分 } \text{Var}(\text{light} * \text{target}) = 0.700$$

读者可以自己解出随机因素的主效应的方差成分值为 0.759。

下面列出不同分析方法的方差成分分析结果，可以看出，不同方法输出结果在数值上稍有差别，但趋势一致，结论一致。

表 9-74 是方差成分表，可以看出，设计的模型包括除了 light\*device 外所有可能的与亮度变量 light 有关的效应项，其方差估计值的总和为：

$$\begin{aligned} &\text{Var}(\text{light})+\text{Var}(\text{light}*\text{target})+\text{Var}(\text{light}*\text{target}*\text{device}) \\ &=0.759+0.700+4.012=5.471 \end{aligned}$$

light 的各阶效应的总方差估计值/ (light 的各阶效应的总方差估计值：+Var(Error))=5.471/(5.471+0.733)=88.19%。

亮度的效应解释了随机效应的 88.19%，误差效应仅解释了随机效应的 11.91%。说明了亮度对随机效应的贡献相当多。在该项心理测试实验中是不可忽视的。而在该表中还可以看出三维交互项所解释的方差是 4.012，在亮度的所有可能的效应中占了 73.3%。因此三维效应又是最值得关注的。

表 9-75(a)ML 方法的方差成分分析结果。表 9-75(b)REML 方法的方差成分分析结果。表 9-75(c)MINQUE 方法的方差成分分析结果。

以上四种方法均说明方差最大来源于亮度、目标、设备的交互效应。亮度因素是不可忽视的，亮度应该在测试中作为测试条件考虑。

表 9-75 不同方法的方差成分分析结果

表 9-74 方差成分表  
(去掉 Light\*device)

Variance Estimates	
Component	Estimate
Var(light)	.759
Var(light * target)	.700
Var(light * target * device)	4.012
Var(Error)	.733

Dependent Variable: score  
Method: ANOVA (Type III Sum of Squares)

Variance Estimates	
Component	Estimate
Var(light)	.348
Var(light * target)	.000 <sup>a</sup>
Var(light * device)	.000 <sup>a</sup>
Var(light * target * device)	3.353
Var(Error)	.733

Dependent Variable: score  
Method: Maximum Likelihood Estimation  
a. This estimate is set to zero because it is redundant.

(a)

Variance Estimates	
Component	Estimate
Var(light)	.759
Var(light * target)	.700
Var(light * device)	.000 <sup>a</sup>
Var(light * target * device)	4.012
Var(Error)	.733

Dependent Variable: score  
Method: Restricted Maximum Likelihood Estimation  
a. This estimate is set to zero because it is redundant.

(b)

Variance Estimates	
Component	Estimate
Var(light)	.759
Var(light * target)	.700
Var(light * target * device)	4.012
Var(Error)	.733

Dependent Variable: score  
Method: Minimum Norm Quadratic Unbiased Estimation (Weight = 1 for Random Effects and Residual)

(c)

习 题 9

- 1. 简述方差分析的基本思想。用表达式表示单因素方差分析的偏差平方和分解。
- 2. 方差分析的假定的前提条件有哪些？
- 3. 什么是主效应？什么是交互效应？
- 4. 简述协方差分析的基本思想。
- 5. 对四个服务行业的服务质量进行评价，较高得分表示较高的服务质量。对航空公

司、零售业、旅馆业和汽车制造业进行的评定数据见 data09-12。在显著性水平  $\alpha=0.05$  下，检验 4 种行业质量等级的总体均值是否差异显著？你的结论如何？

6. 数据 data09-13 是 474 个银行职工的数据。试分析银行办事员起始工资是否与职工的性别、民族有关？分析时假定银行办事员起始工资总体为正态分布，并考虑其他因素的影响。

7. 数据 data09-14 是 15 名手术要求基本相同的患者，随机分 3 组，分别在手术中使用三种麻醉诱导方法 A, B, C，在不同时相（诱导前 T0 和 T1、T2、T3、T4）测量的收缩压数据。试进行方差分析。

# 第 10 章 相 关 分 析

## 10.1 相关分析的概念与相关分析过程

### 10.1.1 简单相关分析的概念

#### 1. 两个变量间的简单相关分析

相关分析是研究变量间密切程度的一种常用统计方法。线性相关分析研究两个变量间线性关系的强弱程度和方向。相关系数是描述线性关系强弱程度和方向的统计量，通常用  $r$  表示。

如果一个变量  $y$  可以确切地用另一个变量  $x$  的线性函数表示，这种关系是确切的，则两个变量间的相关系数是 1 或 -1。

一般情况，两个变量的对应关系不具有唯一性。例如身高与体重的关系，相同身高的人会有不同的体重。研究它们之间线性关系的密切程度使用相关分析。

变量  $y$  随着变量  $x$  的增加而增加或随着变量  $x$  的减少而减少，称为变化方向一致。发育阶段的少年，身高越高，体重相对也就越大。这种相关称为正向相关，其相关系数大于 0。如果变量  $y$  随着变量  $x$  的增加而减少，例如吸烟量和吸烟时间与肺功能的关系，变化方向相反。随着吸烟量增加，肺功能下降；随着吸烟时间加长，肺功能下降。这种相关关系称为负相关。相关系数小于 0。相关系数  $r$  没有单位，其值在 -1~1 之间。

正态分布的等间隔测度变量  $x$  与  $y$  间的相关系数采用 Pearson 积矩相关公式计算

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

式中， $\bar{x}$ 、 $\bar{y}$  分别是变量  $x$ 、 $y$  的均值。 $x_i$ 、 $y_i$  分别是变量  $x$ 、 $y$  的第  $i$  个观测值。

#### 2. 非参相关分析

如果数据分布不满足正态分布的条件，应使用 Spearman 和 Kendall 相关分析方法。

(1) Spearman 相关系数是 Pearson 相关系数的非参形式，是根据数据的秩而不是根据实际值计算的。也就是说，先对原始变量的数据排秩，根据各秩使用 Spearman 相关系数公式进行计算。它适合有序数据或不满足正态分布假设的等间隔数据。相关系数值的范围也是在 -1~1 之间，绝对值越大，表明相关性越强。相关系数的符号也表示相关的

方向。这两种相关系数的计算必须对变量值排序。变量  $x$ 、 $y$  之间的 Spearman 相关系数计算公式为

$$\theta = \frac{\sum (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum (R_i - \bar{R})^2 \sum (S_i - \bar{S})^2}}$$

式中,  $R_i$  是第  $i$  个  $x$  值的秩,  $S_i$  是第  $i$  个  $y$  值的秩。 $\bar{R}$ 、 $\bar{S}$  分别是  $R_i$  和  $S_i$  的平均值。

(2) Kendall's tau-b 也是一种对两个有序变量或两个秩变量间的关系程度的测度, 因此也属于一种非参测度。分析时考虑了结点(秩次相同的)的影响。Kendall's tau-b 计算公式如下

$$\tau = \frac{\sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}$$

其中,  $\text{sgn}(z) = \begin{cases} 1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$

$$T_0 = n(n-2)/2; \quad T_1 = \sum t_i(t_i-1)/2; \quad T_2 = \sum u_i(u_i-1)/2$$

$t_i$  (或  $u_i$ ) 是  $x$  (或  $y$ ) 的第  $i$  组结点  $x$  (或  $y$ ) 值的数目,  $n$  为观测量数。

两个或若干变量之间或两组观测量之间的关系, 有时也可以用相似性或不相似性来描述。相似性测度用大数值表示很相似, 较小的数值表明相似性小。不相似性使用距离或不相似性来描述, 大值表示相差甚远, 有关内容请参见 10.4 节。

### 3. 关于相关系数统计意义的检验

由于我们通常是通过抽样方法, 利用样本研究总体的特性。由于抽样误差的存在, 样本中两个变量间相关系数不为 0, 不能说明总体中这两个变量间的相关系数不是 0, 因此必须经过检验。检验的零假设是: 总体中两个变量间的相关系数为 0。SPSS 的相关分析过程给出了该假设成立的概率, Pearson 和 Spearman 相关系数假设检验  $t$  值计算公式

$$t = \frac{\sqrt{n-2} \cdot r}{\sqrt{1-r^2}}$$

式中,  $r$  是相关系数,  $n$  是样本观测量数,  $n-2$  是自由度。当  $t > t_{0.05(n-2)}$  时,  $p < 0.05$  拒绝原假设; 否则不足以在这个检验中拒绝相关系数为 0 的原假设。

在 SPSS 的相关分析过程的输出中只给出相关系数和假设成立的概率  $p$  值。

## 10.1.2 相关分析过程

在 Analyze 下拉菜单中的 Correlate 命令项具有 3 个相关分析功能命令, 见图 10-1。

1. Bivariate 命令项调用 Correlations 过程和 Nonpar Corr 过程, 按指定项显示变量的描述统计量。计算指定的两个变量间的相关系数, 可以选择计算 Pearson 相关系数、

Spearman 相关系数和 Kendall's tau-b 相关系数。同时对相关系数进行检验。检验的零假设是：总体中两个变量间的线性相关系数为 0。可以对检验进行单尾或双尾的选择。给出相关系数为 0 的概率。

2. Partial 命令项调用 Partial Corr 过程，计算两个变量间在控制了其他变量的影响下的相关系数。可以选择单尾或双尾显著性检验。检验的零假设是：总体中两个变量间偏相关系数为 0。还可以要求计算其他描述统计量。

3. Distance 命令项调用 Proximities 过程，对变量或观测测量进行相似性或不相似性测度。因此分析的变量可以是连续变量、表示频数分布的变量，对某些测度还可以适用于二值变量。可以对原始数据和计算出的距离数据进行标准化。

如果为达到预测目的，研究自变量的变动对因变量的影响程度，根据已知自变量的变化来估计因变量的变化情况，必须使用回归分析。

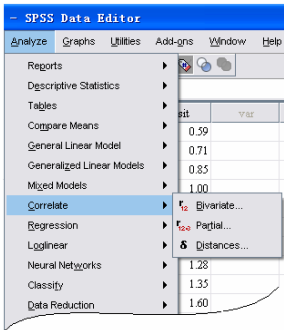


图 10-1 相关分析菜单

## 10.2 两个变量间的相关分析

本节介绍两变量间的相关，包括两个连续变量间的相关和两个等级变量间的秩相关。这两种相关使用同一个命令项 Bivariate 调用。在对话框中，通过选择不同的分析方法调用不同的分析过程。选择哪一种分析方法要看具体的数据类型。

### 10.2.1 两变量间相关分析过程

在进行相关分析之前，应该使用 Graphs 菜单中的 Scatter 命令作散点图，进行初步观察，确认两个变量间有相关趋势，再按下列步骤进行相关分析。

#### 1. 选择分析变量

按 Analyze→Correlate→Bivariate 顺序单击菜单项，展开双变量相关分析主对话框，如图 10-2 所示。在左面的变量表中选择两个以上变量送入 Variables 框中。

2. Correlation Coefficients 栏中列出相关分析类型，有 3 个选项：

(1) Pearson，皮尔逊相关，系统默认的相关分析方法。只有正态分布的等间隔测度的变量才使用这种相关分析。

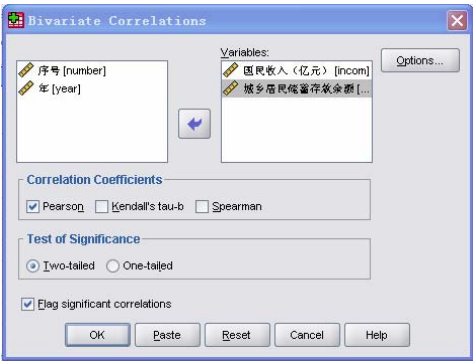


图 10-2 二元变量相关分析主对话框

(2) Spearman, 斯皮尔曼相关, 调用 Nonpar Corr 过程计算斯皮尔曼秩相关系数。

(3) Kendall's tau-b, 肯德尔 $\tau$ -b, 调用 Nonpar Corr 过程考虑结点的影响, 计算分类变量间的秩相关。

如果参与分析的变量是连续变量, 选择 Kendall's tua-b 或 Spearman 相关, 则系统自动对连续变量的值先求秩, 再计算其秩分数间的相关系数。

### 3. Test of Significance 栏, 显著性检验选项

(1) Two-tailed, 双尾 T 检验, 系统默认的检验方式, 当事先不知道相关方向 (正相关还是负相关) 时选择此项。

(2) One-tailed, 单尾 T 检验, 如果事先知道相关方向可以选择此项。

检验针对的零假设是: 总体中两个变量不相关。显示假设成立的概率水平。

4. Flag significant correlations, 要求在输出结果中, 相关系数右上方使用 “\*” 表示显著性水平为 5%, 用 “\*\*” 表示其显著性水平为 1%。

### 5. Options 对话框中的选择项

在主对话框中单击 Options 按钮, 展开如图 10-3 所示对话框。

#### (1) 统计量选项

只有在主对话框中选择了 Pearson 相关分析方法时才可以选择这两个选项。

① Means and standard deviations, 输出均值与标准差。

② Cross-product deviations and covariances, 输出叉积离差矩阵和协方差矩阵。

(2) 缺失值处理方法选项。在 Missing Values 栏中:

① Exclude cases pairwise, 仅剔除正在参与计算的两个变量值是缺失值的观测量。这样, 有可能在计算出的相关系数矩阵中, 相关系数是根据不同数量的观测量计算出来的。选择此项, 可以最大限度地使用取得的观测数据。

② Exclude cases listwise, 剔除在主对话框 Variables 框中列出的变量带有缺失值的所有观测量。输出的相关矩阵中, 每个相关系数都是依据相同数量的观测量计算出来的。

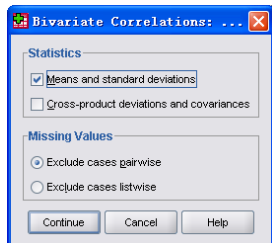


图 10-3 输出选项对话框

## 10.2.2 两个变量间相关分析实例

【例 1】使用默认选项进行简单相关分析的例题

(1) 数据文件 data10-01。数据说明: 以 1962~1988 年安徽省国民收入与城乡居民储蓄存款余额两个变量间的线性相关分析为例, 说明使用系统默认值进行连续变量相关分析的方法。数据来源于中国现场统计研究会主办的《数理统计与管理》1990 年第 5 期。数据编号 data010-01。变量包括: income (国民收入 (亿元)), deposit (城乡居民储蓄存款余额), number (序号), year (年份)。

(2) 操作说明: 在源变量栏中选择分析变量 deposit 和 income, 单击向右箭头按钮,



将选择的变量移至 Variables 矩形框中。其余使用系统默认选项。单击 OK 按钮提交系统执行。单击 Paste 按钮得到下列程序。

(3) 程序及其解释：

CORRELATIONS	①
/VARIABLES=deposit income	②
/PRINT=TWOTAIL SIG	③
/MISSING=PAIRWISE .	④

其中

- ① CORRELATIONS 命令调用相关分析过程。
  - ② VARIABLES 子命令指定要计算相关系数的变量表。
  - ③ PRINT 子命令要求输出双尾 T 检验的显著性概率。
  - ④ MISSING 子命令要求在进行相关分析时，成对剔除带有缺失值的观测量。
- (4) 输出结果见表 10-1。

表 10-1 安徽省国民收入与城乡居民存款储蓄余额的相关分析

Correlations			
		城乡居民储蓄存款余额	国民收入（亿元）
城乡居民储蓄存款余额	Pearson Correlation	1	.976**
	Sig. (2-tailed)		.000
	N	27	27
国民收入（亿元）	Pearson Correlation	.976**	1
	Sig. (2-tailed)	.000	
	N	27	27

\*\* . Correlation is significant at the 0.01 level (2-tailed).

表 10-1 是安徽省国民收入变量 income 和城乡居民存款余额 deposit 之间的相关系数矩阵，在变量行与变量列的交叉处纵向显示了 3 个数值。

第一行中的数值是行变量与列变量的相关系数矩阵。行、列变量相同，其相关系数为 1。变量 income 与 deposit 之间的相关系数为 0.976。

第二行中的数值是使相关系数为 0 的假设检验成立的概率，结果均小于 0.001。

第三行中的数值是参与该相关系数计算的观测量数目，均为 27。

注释行说明标有 “\*\*” 的相关系数的显著性概率水平为 0.01。显然，国民收入与存款余额之间是高度相关的。

【例 2】生成矩形相关矩阵的简单相关例题

(1) 数据 data10-02.说明：本例题为一组银行雇员数据。分析的目的是要观察 salbegin（起始工资）和 salary（现工资）与雇员本人各方面条件的关系。变量有：salary（当前工资）、age（年龄）、jobtime（以月为本单位的工作时间）、prevexp（以月为本单位的以前工作经历）。

(2) 操作步骤：

- ① 读取数据文件，按 Analyze→Correlate→Bivariate 顺序单击菜单项，展开对话框。

② 在源变量框中选择 jobtime、prevexp、age、salary、salbegin 送入 Variables 栏作为分析变量。

③ 主对话框中的选项：

- 分析方法选择 Person 相关。
- 显著性检验选择 Two-tailed 双尾 T 检验。
- 选中 Flag significant correlations 复选项。

④ 在 Options 窗口中指定选项：

- 要求计算的统计量选择 Means and standard deviations 要求计算均值与标准差。
- 缺失值处理方法选择 Exclude cases pairwise 成对剔除带有缺失值的观测量。

返回主对话框，单击 OK 按钮提交运行。

⑤ 运行程序语句

CORRELATIONS

①

/VARIABLES= jobtime prevexp age salary salbegin

②

/PRINT=TWOTAIL NOSIG

③

/STATISTICS DESCRIPTIVES

④

/MISSING=PAIRWISE .

⑤

生成程序后进行修改，由于只需要变量 salary、salbegin 与其他各变量的相关性，因此在此第②语句 salary、salbegin 与其他变量之间增加 with，以便使结果更加清晰。

即

/VARIABLES= salary with salbegin jobtime age prevexp.

(3) 程序运行结果见表10-2和表10-3。

表10-2 分析变量的描述统计量

Descriptive Statistics			
	Mean	Std. Deviation	N
当前工资	\$34,419.57	\$17,075.661	474
起始工资	\$17,016.09	\$7,870.638	474
受雇月数	81.11	10.061	474
年龄	47.14	11.775	473
过去经验(月)	95.86	104.586	474

表10-3 Pearson相关系数矩阵

Correlations				
	起始工资	受雇月数	年龄	过去经验(月)
当前工资	Pearson Correlation .880**	.084	-.144*	-.097*
	Sig. (2-tailed) .000	.067	.002	.034
	N 474	474	473	474

\*\* . Correlation is significant at the 0.01 level (2-tailed).

\* . Correlation is significant at the 0.05 level (2-tailed).

(4) 结果分析

表10-2是相关分析变量的描述统计量。可以看出工资平均值当前工资比起始工资高，而且现在工资标准差比起始工资标准差大了，说明现在工资差别大了。

表 10-3 表在行变量与列变量的交叉点单元格上，自上至下 3 个统计量分别为：

Pearson Correlation 皮尔逊相关系数。表中的相关系数均很小。

表中 Sig. (2-tailed) 是对于相关系数为 0 的假设的双尾 T 检验结果，是 t 值大于其临界值的概率  $p$ 。N 为参与相关系数计算的有效观测量数。

很明显，当前工资与起始工资相关系数最大 0.88，不相关的概率小于 0.001；与受雇

月数几乎无关，与年龄成负相关，不相关的概率很低，为0.01。年龄越大，工资有越低的趋势。与以前工作经历相关系数更低，只有-0.097；但不相关的概率为0.034，小于0.05，说明当前工资考虑了工作经历，但是工作经历占的比重不大。

【例 3】秩相关实例。

(1) 数据 data10-02。说明：以上对雇员的工资的分析并不严格，因为虽然参与分析的变量均为尺度（连续）变量，但没有做各变量是否符合正态分布的检验。下面使用秩相关分析方法分析各雇员的 jobcat（职务等级）、educ（受教育程度）与 salary（当前工资）、salbegin（起始工资）间的关系。educ 数值数小于 24（Options 对话框中定义的），因此标为 Ordinal 属于有序分类变量。

(2) 操作：重新启动 Bivariate Correlate，移入 Variable 框中的变量有 salbegin、salary、educ、jobcat。分析方法选择 Kendall's tau-b 秩相关。选择 Two-tailed 双尾 T 检验。选中 Flag significant correlations 复选项。缺失值处理方法选择 Exclude cases pairwise 成对剔除带有缺失值的观测量。

(3) 运行程序语句：在主对话框，单击 Paste 按钮，在语句窗口中生成如下程序。

```
NONPAR CORR /VARIABLES= salary salbegin jobcat educ
/PRINT=KENDALL TWOTAIL NOSIG
/MISSING=PAIRWISE .
```

生成程序后对语句进行修改，修改 VARIABLES 语句成如下形式：

```
/VARIABLES= salary with salbegin jobcat educ
```

(4) 输出结果见表 10-4。

表 10-4 非参相关矩阵

(5) 输出结果分析  
表 10-4 中现工资与起始工资秩相关系数较高为 0.656，和受教育水平的秩相关系数为 0.554，不相关的概率几乎为 0，与工作分类的秩相关系数为 0.530，不相关的概率几乎为 0。结合上例结果可以看出，起始工资与受雇时间、过去工作经验相关系数都很小，而不相关的概率均大于 5%。可以说，起始工资和现工资仅、受教育水平和职务有关。

Correlations					
	当前工资	起始工资	工作分类	受教育水平 (年)	
Kendall's tau_b		.656**	.530**	.554**	
	Correlation Coefficient				
	Sig. (2-tailed)	.000	.000	.000	
	N	474	474	474	

\*\* . Correlation is significant at the 0.01 level (2-tailed).

读者可以根据自己的经验分析这样的工资结构是否合理，是否有利于调动职工的积极性，从而有利于企业的发展。

【例 4】两个等级变量相关间秩相关实例。

数据 data10-03 为某次全国武术女子比赛前 10 名运动员长拳和长兵器两项得分数据，要求分析这两项得分是否存在线性关系。变量 score1、score2 分别为长拳和长兵器两项得分，ranking 变量为名次。

(1) 读取数据文件 data10-03。

(2) 按 Analyze→Correlate→Bivariate 顺序单击菜单项，展开两变量相关分析主对话

框。在主对话框中的选项：分析变量：score1、score2；分析方法：Kendall's tau-b、Spearman；显著性检验类型：One-tailed 单尾 T 检验；选择 Flag significant correlations。

(3) Options 对话框中的选择分析中成对剔除带有缺失值的观测量。改变 Varilbles 子命令，加 With 在两个变量之间。

(4) 运行的程序如下：

```
NONPAR CORR  /VARIABLES=score1 with score2
              /PRINT=BOTH ONETAIL NOSIG  /MISSING=PAIRWISE .
```

(5) 输出结果见表 10-5。

从表中可以看到，Kendall's tau-b 相关系数为 0.543，不相关的假设成立的概率为 Sig=0.027，由于采用的是单尾检验，所以不相关的概率为 0.054 大于 0.05；从 Spearman 相关系数是 0.610，假设成立的概率为  $p=0.060$ ，大于 0.05。可以得出结论，两种分析方法等级相关系数统计意义上不显著（5%水平），不能拒绝两项比赛得分间不存在相关关系的假设。

表 10-5 Kendall's tau-b 与 Spearman 相关系数

Correlations			
Kendall's tau_b	长拳得分	Correlation Coefficient	.543 <sup>*</sup>
		Sig. (1-tailed)	.027
		N	10
Spearman's rho	长拳得分	Correlation Coefficient	.610 <sup>*</sup>
		Sig. (1-tailed)	.030
		N	10

\*. Correlation is significant at the 0.05 level (1-tailed).

对于非等间隔测度的连续变量，因为分布不明，可以使用等级相关分析，如上一个例题也可以使用 Pearson 相关分析。对于完全等间隔的离散变量，必须使用等级相关分析相关性。

### 10.2.3 两个变量相关分析的过程语句

两变量间相关分析包括对等间隔测度变量的 Pearson 相关分析和等级相关分析 Spearman、Kendall's tau-b。分别使用不同的相关分析过程。

#### 1. Pearson 相关分析的过程语句

Pearson 相关分析使用 CORRELATIONS 命令

```
CORRELATIONS [VARIABLES=] varlist [WITH varlist] [/varlist...]
              [/MISSING={PAIRWISE**} [{INCLUDE}{LISTWISE}{EXCLUDE}]]
              [/PRINT={TWO TAIL**} {SIG**}{ONETAIL}{NOSIG}]
              [/MATRIX=OUT({* }{file})]
              [/STATISTICS={DESCRIPTIVES} [XPROD] [ALL]]
```

#### (1) CORRELATIONS 命令语句

进行 Pearson 相关分析调用 CORRELATIONS 过程语句。CORRELATIONS 是命令语句关键字，关键字后面必须至少出现两个变量组成的变量表。变量表有以下几种形式：

① varlist，直接书写一系列参与分析的变量名。分析结果将给出这些变量两两间的

相关系数方阵。对每一个相关系数给出参与计算的观测量数和相对于相关系数为 0 的假设检验的显著性概率。

② **varlist WITH varlist**, 即使用 **WITH** 连接的两个变量表。分析结果将给出 **WITH** 前变量表中的变量与 **WITH** 后面的变量表中的变量之间的相关系数矩阵, 如表 10-4。

可以在一切操作、选择均完成后, 单击 **Paste** 按钮, 在 **Syntax** 窗口中生成命令程序。将 **CORRELATIONS** 语句修改成使用 **WITH** 连接两个变量表的形式, 然后在该窗口中单击运行按钮, 提交运行, 即可得到更简洁的矩阵。

③ **VARIABLES= varlist**, 效果同①; **VARIABLES= varlist WITH varlist**, 效果同②。

④ **VARIABLES=varlist / varlist**, 效果同②; **varlist / varlist**, 效果同②。

## (2) **MISSING** 子命令

该子命令指定处理缺失值的方法。让缺失值参与分析的 **INCLUDE** 选项是不可取的, 一般不用。另外两个选项可以任取其一:

① **PAIRWISE**, 成对剔除带有缺失值的观测量, 即只在计算两个变量的相关系数时, 剔除这两个变量为缺失值的观测量, 这样, 相关系数矩阵中的各相关系数可能是根据不同数目的观测计算出来的。此为系统默认的处理方法。

② **LISTWISE VARIABLES**, 变量表中任何一个变量带有缺失值, 对应的观测量要从所有相关系数计算中剔除。合法观测量数目对所有的分析都相同。

③ **INCLUDE**, 把读者定义的缺失值当做合法值处理。

(3) **PRINT** 子命令指定是否进行显著性检验和使用单尾还是双尾显著性检验。

① **SIG** 与 **NOSIG**, 指定是否按显著性检验结果标出不相关的概率水平在 5% 和 1% 的相关系数。**SIG** 是系统默认的, 系统默认为 1%  $< p < 5\%$  标 “\*”,  $p < 0.01$  标 “\*\*”。

② **ONETAIL**, 指定进行单尾显著性检验; **TWOTAIL**, 指定进行双尾显著性检验。

(4) **FORMAT** 子命令指定输出相关系数的方式:

① **MATRIX**, 指定以矩阵方式输出相关分析结果。此为系统默认方式。

② **SERIAL**, 指定以行方式输出相关分析结果。

(5) **STATISTICS** 子命令指定除相关分析的结果外要求输出的统计量。

① **DESCRIPTIVES**, 要求输出描述统计量, 包括各分析变量的均值和标准差。

② **XPROD**, 要求计算并输出叉积离差矩阵和协方差矩阵。

③ **ALL**, 要求输出描述统计量、叉积离差矩阵和协方差矩阵。

## 2. 非参数相关分析的命令语句

非参相关分析调用 **NONPAR CORR** 过程。该过程的命令语句如下:

```
NONPAR CORR [VARIABLES=] varlist [WITH varlist] [/varlist...]

[/PRINT={TWOTAIL**}{ONETAIL}{SIG**}{NOSIG}{SPEARMAN**}{KENDALL}
{BOTH}]

[/SAMPLE]
```

```
[/MISSING={PAIRWISE**} {LISTWISE} {INCLUDE}]
```

```
[/MATRIX=OUT({*} {file})]
```

### (1) NONPAR CORR 命令语句

该命令语句除关键字与 **CORRELATIONS** 不同外，变量表书写方式及其对应的输出均与 **CORRELATIONS** 相同。

### (2) 子命令

**NONPAR CORR** 要求使用的子命令基本与 **CORRELATIONS** 的子命令相同。其不同之处只有 **PRINT** 子命令。在该子命令中除了可以指定是否进行显著性检验和单尾还是双尾检验外，还可以指定分析方法：

① **SPEARMAN**，指定计算 Spearman 相关系数。此为系统默认的分析方法。

② **KENDALL**，指定进行 Kendall's tau-b 相关分析。

③ **BOTH**，指定使用以上两种方法进行等级相关分析。

其余子命令的用法请参见本节关于 Pearson 相关分析过程语句的有关内容。

## 10.2.4 关于相关矩阵

在 **CORRELATIONS** 过程和 **NONPAR CORR** 过程中都有 **MATRIX** 子命令，可以将计算出的相关矩阵写入一个数据文件，以便做其他分析时使用。

### 1. CORRELATION 过程中的 MATRIX 子命令

**MATRIX** 把矩阵数据写到一个数据文件。数据包括每个变量的均值、标准差和用于计算相关系数的观测量数目。**PARTIAL CORR**，**REGRESSION**，**FACTOR** 和 **CLUSTER** 等分析过程可以读取由 **CORRELATIONS** 生成的矩阵数据。

### 2. 使用 MATRIX 子命令的方法和注意事项

(1) **CORRELATIONS** 不能把矩形矩阵写入数据文件（用关键字 **WITH** 指定的矩阵）。

(2) 如果在 **CORRELATIONS** 变量表中指定的变量表不只一个，仅最后一个不使用 **WITH** 的变量表生成的相关矩阵写入到矩阵数据文件。

(3) 关键字 **OUT** 指定把矩阵写入哪个文件，文件名必须放在引号中。

(4) 原始数据不包括在矩阵文件中。如果矩阵文件成为工作数据文件，原始数据不会出现。

(5) **OUT (filename)** 把矩阵数据写到文件，在括号中指定文件名或“\*”。如果指定的是一个包括路径的文件名，文件按路径保存到磁盘，可以在任何时候再对其进行处理。如果使用星号，矩阵代替当前工作数据文件，可以使用 **SAVE** 命令将其保存到磁盘。

### 3. 矩阵数据文件的格式

矩阵数据文件有两个特殊的变量，是 **ROWTYPE\_** 和 **VARNAME\_**。

(1) 变量 **ROWTYPE\_** 是一个短字符串变量，其值有 **MEAN** 均值、**STDDEV** 标准差、**N** 参与计算的有效观测量数和 **CORR** 相关系数。

(2) VARNAME\_也是一个短字符串变量，它的值是相关矩阵中的变量名，当 ROWTYPE\_值是 CORR 时，VARNAME\_ 给出相关矩阵的行变量名。

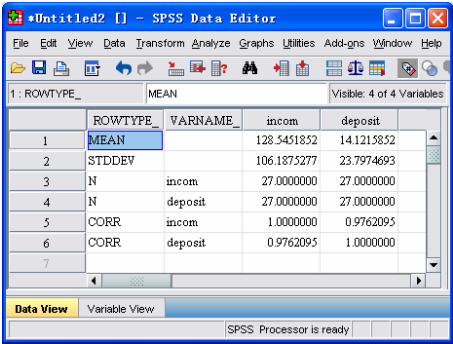


图 10-4 在工作数据窗口中生成的矩阵文

子命令使用方法和注意事项相同，矩阵数据文件也有两个特殊的变量，是 ROWTYPE\_、VARNAME\_。需要说明的是变量 ROWTYPE\_是短字符串变量，对 Spearman 的相关系数来说，其值有 N 和 RHO。对 Kendall’s 系数，其值为 N 和 TAUB。

10.2.5 建立相关矩阵数据文件

如果已知数据是相关矩阵，需要建立的数据文件就不能使用前面章节介绍的方法。建立矩阵数据文件的方法有两种，一是直接把矩阵数据按照 SPSS 的规则输入到数据窗口中；二是使用命令 MATRIX DATA 语句编写 SPSS 程序，运行程序的结果在数据窗口中建立相关矩阵格式的数据，存盘后形成数据文件。

相关矩阵数据文件由可以处理矩阵资料的 SPSS 过程读取，这样的数据文件除包括矩阵数据外，还可以包括各变量的均值、标准差等，见图 10-4。

1. 在数据窗口中直接输入矩阵数据

在数据窗口中定义下列变量：

(1) 变量 ROWTYPE\_，变量类型为短字符串型（变量长度≤8 个字符），它的值为各观测量的统计学名称，其值必须使用规范的统计量名称。例如常用的有：

- ① MEAN 表示各变量均值。
- ② STDDEV（或 SD）表示各变量的标准差。
- ③ N（或 N\_VECTOR）表示各变量的非缺失值的数目。
- ④ CORR 表示相关矩阵的相关系数。
- ⑤ COV 表示协方差阵的系数。
- ⑥ DEF 表示自由度。
- ⑦ MSE 表示误差的均方。

图 10-4 是本章第一个例题，1962~1988 年安徽省国民收入与城乡居民储蓄存款余额两个变量间的线性相关分析，运行的程序加了一个子命令 /MATRIX=OUT(\*)的运行结果。

4. NONPAR CORR 过程中的 MATRIX 子命令

NONPAR CORR 过程中的 MATRIX 子命令与 CORRELATIONS 过程中的 MATRIX

(2) 变量 `VARNAME_`，变量类型为短字符串型（变量长度 $\leq 8$  个字符），它的值为参与分析的（矩阵中涉及的）各变量的变量名。

(3) 分析变量，按前面章节介绍的方法定义分析变量，只不过输入数据时是按 `ROWTYPE_` 定义的值输入相应的数据。例如 `data010-03` 数据文件中的分析变量为 `hgrow`、`temp`、`rain`、`hsun`、`humi`。

## 2. 使用 SPSS 程序语句生成矩阵数据文件

(1) 以相关矩阵的生成程序为例，最简单的生成矩阵数据文件的 SPSS 程序如下：

MATRIX DATA	①
VARIABLES=ROWTYPE_ SAVINGS POP15 POP75 INCOME GROWTH.	②
BEGIN DATA	③
MEAN 9.6710 35.0896 2.2930 1106.7784 3.7576	④
STDDEV 4.4804 9.1517 1.2907 990.8511 2.8699	⑤
N 50 50 50 50 50	⑥
CORR 1	⑦
CORR -.4555 1	⑦
CORR .3165 -.9085 1	⑦
CORR .2203 -.7562 .7870 1	⑦
CORR .3048 -.0478 .0253 -.1295 1	⑦
END DATA.	⑧

### (2) 语句规则

① 第一个语句必须是 `MATRIX DATA` 语句。`MATRIX DATA` 是关键字，必须原样照写。

② 关键字后面紧跟 `VARIABLES=` 变量名表。按各变量在矩阵中出现的自左至右的顺序列出。变量名中必须有一个 `ROWTYPE_` 变量，变量名用空格隔开，结尾加半角圆点作为结束符号。

③ 在数据表列之前使用 `BEGIN DATA` 语句表明下一行是数据。

④～⑦ 数据表列，紧跟 `BEGIN DATA` 语句后面。每个数据行的第一个值是 `VARIABLES` 语句中第一个变量 `ROWTYPE_` 的值，必须是 SPSS 语句允许的规范的关键字值，可以是 `N`、`MEAN`、`STDDEV`（或 `SD`）`CORR`、`COV`、`MSE` 等。

每个数据行的第二个、第三个……值对应于 `VARIABLES` 语句中第二个、第三个……变量的统计量，是什么统计量由所在行的第一个，也就是 `ROWTYPE_` 变量的值决定。各数据行的值的顺序应该与 `VARIABLES=` 后面的变量顺序一致。下面具体说明之。

④ 这行是各变量的均值。`MEAN` 是变量 `ROWTYPE_` 的值，后面的 9.6710、35.0896、2.2930、1106.7784、3.7576 分别是变量 `SAVINGS`、`POP15`、`POP75`、`INCOME`、`GROWTH` 的均值。



⑤ 这行是各变量的标准差。STDDEV 是变量 ROWTYPE\_ 的值，后面的 4.4804、9.1517、1.2907、990.8511、2.8699 分别是变量 SAVINGS、POP15、POP75、INCOME、GROWTH 的标准差。

⑥ 这行是各变量参与计算相关矩阵的观测量数。N 为变量 ROWTYPE\_ 的值。

⑦ 所有标有⑦的数据行组成相关矩阵。每行第一个值都是 CORR，表明后面的都是相关系数。CORR 是变量 ROWTYPE\_ 的值。每个数值在标有 CORR 的数据行中出现的行数  $m$  和列数  $n$  就是 VARIABLES 语句 ROWTYPE\_ 变量后面出现的第  $m$  个变量和第  $n$  个变量的相关系数。参照图 10-5 可以知道数据输入的顺序。

⑧ 以 END DATA 结束数据表列。

**注意，最后一个语句使用半角圆点结束。**

与直接在数据窗口中定义变量、输入数据的方法不同，VARNAME\_ 变量不要出现在变量表中，而是程序执行时自动生成的。

程序输入完毕单击执行图标按钮（一个向右的箭头按钮），执行的结果如图 10-5 所示。可以看出，最简单的程序生成的矩阵文件是方阵。

**注意，过程语句前面不要加任何空格，否则不能生成正确的矩阵数据文件。**

MATRIX DATA 过程语句还有很多子命令。例如 FORMAT 子命令，使用这个子命令的选项可以生成上半三角、下半三角或全格式的矩阵。使用 CONTENTS 子命令，可以节省数据的输入工作量等。读者需要了解这些子命令及其使用规则，可以使用 HELP 功能中的 Syntax Guide 命令，在小菜单中选择 Base 命令，调用阅读器 Acrobat Reader，打开 Spssbase.pdf 文件，在左侧的菜单中选择 MATRIX DATA 可以看到全部有关矩阵数据生成过程的语句解释和使用规则。

	ROWTYPE_	VARNAME_	SAVINGS	POP15	POP75	INCOME	GROWTH
1	N		50.0000	50.0000	50.0000	50.0000	50.0000
2	MEAN		9.6710	35.0896	2.2930	1106.7784	3.7576
3	STDDEV		4.4804	9.1517	1.2907	990.8511	2.8699
4	CORR	SAVINGS	1.0000	.	0.3165	0.2203	0.3048
5	CORR	POP15	.	1.0000	.	.	.
6	CORR	POP75	0.3165	.	1.0000	0.7870	0.0253
7	CORR	INCOME	0.2203	.	0.7870	1.0000	.
8	CORR	GROWTH	0.3048	.	0.0253	.	1.0000
9	.	.	.	.	.	.	.

图 10-5 程序生成的矩阵数据文件

## 10.3 偏相关分析

### 10.3.1 偏相关分析的概念

#### 1. 偏相关分析

简单相关分析计算两个变量间的相关系数,分析两个变量间线性关系的程度和方向。往往因为第三个变量的作用,使相关系数不能真正反映两个变量间的线性程度。例如身高、体重与肺活量之间的关系。如果使用 Pearson 相关分析计算其相关系数,可以得出肺活量与身高和体重均存在较强的线性关系。但实际上,如果对体重相同的人,分析身高和肺活量。是否身高值越大,肺活量越大呢?结论是否定的。正是因为身高与体重有着线性关系,体重与肺活量存在线性关系,因此,得出身高与肺活量之间存在着较强的线性关系的错误结论。偏相关分析的任务就是在研究两个变量之间的线性相关关系时控制可能对其产生影响的变量。分析身高与肺活量之间的相关性,就要控制体重在相关分析中的影响。实际生活中有许多这样的关系,例如可以控制年龄和工作经验两个变量的影响,估计工资收入与受教育程度之间的相关关系。可以在控制销售能力与各种其他经济指标的情况下,研究销售量与广告费用之间的关系等。

#### 2. 偏相关系数的计算

控制了变量  $z$ , 变量  $x$ 、 $y$  之间的偏相关和控制了两个变量  $z_1$ 、 $z_2$ , 变量  $x$ 、 $y$  之间的偏相关系数计算公式分别为下面两个公式

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}; \quad r_{xy,z_1z_2} = \frac{r_{xy,z_1} - r_{xz_2,z_1}r_{yz_2,z_1}}{\sqrt{(1-r_{xz_2,z_1}^2)(1-r_{yz_2,z_1}^2)}}$$

第一个公式中的  $r_{xy,z}$  是控制了  $z$  的条件下,  $x$ 、 $y$  之间的偏相关系数。 $r_{xy}$  是变量  $x$ 、 $y$  间的简单相关系数或称零阶相关系数。 $r_{xz}$ 、 $r_{yz}$  分别是变量  $x$ 、 $z$  间的和变量  $y$ 、 $z$  间的简单相关系数,以此类推。

#### 3. 偏相关系数的检验

在利用样本研究总体的特性时,由于抽样误差的存在,样本中控制了其他变量的影响,两个变量间偏相关系数不为 0,不能说明总体中这两个变量间的偏相关系数不是 0,因此必须进行检验。检验的零假设:总体中两个变量间的偏相关系数为 0。使用 T 检验方法,公式如下

$$t = \frac{\sqrt{n-k-2} \cdot r}{\sqrt{1-r^2}}$$

这是对 Pearson 偏相关系数假设检验的  $t$  统计量的计算公式,其中,  $r$  是相应的偏相关系数,  $n$  是观测量数,  $k$  是控制变量的数目,  $n-k-2$  是自由度。当  $t > t_{0.05(n-k-2)}$  时,  $p < 0.05$

拒绝原假设，否则不足以在这个检验中拒绝变量间偏相关系数为 0 的零假设。

在 SPSS 的偏相关分析过程的输出中只给出偏相关系数和假设成立的概率  $p$  值。

### 10.3.2 偏相关分析过程

1. 按 Analyze→Correlate→Partial 顺序单击菜单项，展开如图 10-6 所示的偏相关分析主对话框。

2. 从左面的变量表中选择分析变量送入 Variables 矩形框中。选择控制变量送入 Controlling for 框。

3. 在 Test of Significance 栏中选择假设检验类型，有两个选项：

(1) Two-tailed，双尾检验，用于有正负相关两种可能的情况，是系统默认方式。

(2) One-tailed，单尾检验，用于只可能是正向或只可能是负向相关的情况。

4. 是否显示实际的显著性水平

选择 Display actual significance level，在显示相关系数的同时，显示实际的显著性概率。不选择此项，其显著性概率使用星号“\*”代替，表示其显著性概率在 5%~1%之间。

“\*\*”表示其显著性概率小于或等于 1%。

5. Options 窗口中的选项

对话框中单击 Options 按钮，展开如图 10-7 所示对话框。

(1) Statistics 统计量选项

- Means and standard deviations，要求计算并显示各分析变量的均值和标准差。

- Zero-order correlations，要求显示零阶相关矩阵，即 Pearson 相关矩阵。

(2) Missing Values 处理缺失值观测量的选项

- Exclude cases listwise，剔除所有带有缺失值的观测量。系统默认此项。

- Exclude cases pairwise，成对剔除带有缺失值的观测量。

选择完成后，单击 Continue 按钮返回主对话框。单击 OK 按钮提交系统执行。

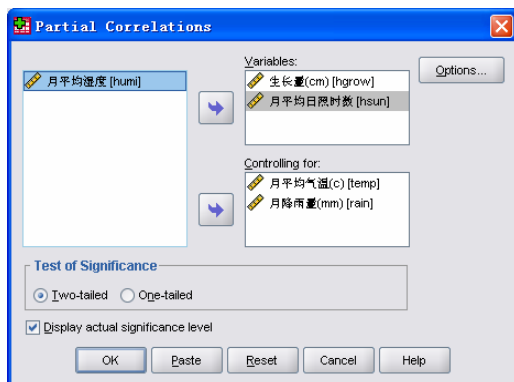


图 10-6 偏相关分析的主对话框

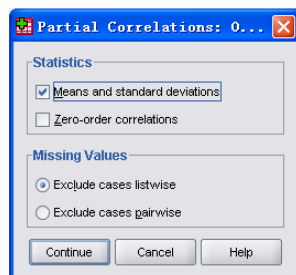


图 10-7 偏相关的选项对话框

### 10.3.3 偏相关分析实例

【例 5】使用四川绵阳地区 3 年生中山柏的数据, 分析月生长量与月平均气温、月降雨量、月平均日照时数、月平均湿度这 4 个气候因素哪个因素有关。数据来源于袁佳祖编著《灰色系统理论》, 数据编号 data10-04。

这 4 个气候因素彼此均有影响, 分析时应分别求生长量与 4 个气候因素分别求偏相关, 在求生长量与一个气候因素的相关时控制其他因素的影响, 然后比较相关系数, 按 4 个气候因素对中山柏生长量的影响的大小排队。

1. 定义变量: month (月份)、hgrow (生长量, cm)、temp (月平均气温,  $^{\circ}\text{C}$ )、rain (月降雨量, mm)、hsun (月平均日照时数)、humi (月平均湿度)。输入数据和求简单相关系数的操作略。

2. 按 Analyze→Correlation→partial 顺序单击菜单项, 启动偏相关分析的主对话框。

3. 指定分析变量和控制变量。

为操作简便, 首先确定第一次分析的变量和控制变量。第一次分析变量是生长量与月平均日照时数 (hgrow 与 hsun), 控制变量是月平均湿度 (humi)、降雨量 (rain)、月平均气温 (temp) 3 个变量。

4. 指定选项

(1) 主对话框中的选项使用系统默认值: 双尾检验, 显示实际的显著性概率。

(2) 在 Options 对话框中不选择任何选择项。假定我们已经对各变量进行过探索分析, 不要各变量的描述统计量。为对比, 我们单写一段计算 Pearson 相关的程序, 以简化相关矩阵 (见第一段程序), 对缺失值的处理使用系统默认的方法。

5. 在主对话框中, 单击 Paste 按钮。在 Syntax 窗口中生成第一次分析的程序:

```
PARTIAL CORR                                ①  
/VARIABLES= hgrow hsun BY humi rain temp    ②  
/SIGNIFICANCE=TWOTAIL                      ③  
/MISSING=LISTWISE                          ④
```

程序解释:

① PARTIAL CORR 语句调用偏相关分析过程。

② VARIABLES 子命令定义分析变量与控制变量。在 BY 前面 hgrow hsun 为要求相关系数的分析变量, BY 后面的是 humi、rain、temp 3 个控制变量。

③ SIGNIFICANCE 子命令要求进行双尾显著性检验。

④ MISSING 子命令要求剔除所有带有缺失值的观测量。

此程序即下面的第 2 段程序 (一个英文据点作为一段程序的结束标志)

6. 复制与修改

(1) 在 Syntax 窗口中, 选择第一次偏相关分析程序, 复制并粘贴三次。

(2) 修改各复制的程序中的 **VARIABLES** 子命令。改变分析变量和控制变量。形成下面的第 3、4、5 段程序。第 1 段程序是求生长量变量与其他气候变量的 **Pearson** 相关系数。作为对比用。

形成以下几个程序段：

```
CORRELATIONS
/VARIABLES=hgrow with hsun humi rain temp
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE .
```

```
PARTIAL CORR
/VARIABLES= hgrow hsun BY humi rain temp
/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE .
```

```
PARTIAL CORR
/VARIABLES= hgrow humi BY hsun rain temp
/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE .
```

```
PARTIAL CORR
/VARIABLES= hgrow rain BY hsun humi temp
/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE .
```

```
PARTIAL CORR
/ hgrow temp BY hsun humi rain
/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE .
```

7. 执行以上各段程序。窗口中显示部分结果见表 10-6、表 10-7。

表 10-6 生长量与各变量间 Pearson 相关分析结果

Correlations					
		月平均日照时 数	月平均湿度	月降雨量(mm)	月平均气温(c)
生长量(cm)	Pearson Correlation	.704 <sup>*</sup>	.374	.709 <sup>**</sup>	.983 <sup>**</sup>
	Sig. (2-tailed)	.011	.232	.010	.000
	N	12	12	12	12

\*. Correlation is significant at the 0.05 level (2-tailed).  
\*\*. Correlation is significant at the 0.01 level (2-tailed).

8. 分析结果解释与结论

从表 10-6 的零阶相关矩阵可以看出，生长量与湿度的相关系数最小，显著性检验结果是不相关的概率为 23%。结论是生长量除与月平均湿度无关外，与其他几个气候因素

均有明显的线性关系。

由于各气候因素的相互影响，例如月平均日照时数与月平均气温高度相关。生长量与各变量间的相关系数并未反映出各变量间的真实情况。因此应该看偏相关的结果。

根据生长量与各气候因素单独的偏相关分析结果，偏相关系数见表 10-7。

表 10-7 偏相关分析结果

Correlations				
Control Variables	Variables		生长量(cm)	月平均日照时数
月平均湿度 & 月降雨量 (mm) & 月平均气温(c)	生长量(cm)	Correlation	1.000	.632
		Significance (2-tailed)	.	.068
		df	0	7
月平均日照时数	月平均湿度	Correlation	.632	1.000
		Significance (2-tailed)	.068	.
		df	7	0

Correlations				
Control Variables		生长量(cm)	月平均湿度	
月平均日照时数 & 月平均湿度 & 月降雨量(mm)	生长量(cm)	Correlation	1.000	.731
		Significance (2-tailed)	.	.025
		df	0	7
月平均湿度	月平均日照时数	Correlation	.731	1.000
		Significance (2-tailed)	.025	.
		df	7	0

Correlations				
Control Variables		生长量(cm)	月降雨量(mm)	
月平均日照时数 & 月平均湿度 & 月降雨量(mm)	生长量(cm)	Correlation	1.000	-.491
		Significance (2-tailed)	.	.180
		df	0	7
月降雨量(mm)	月平均日照时数	Correlation	-.491	1.000
		Significance (2-tailed)	.180	.
		df	7	0

Correlations				
Control Variables		生长量(cm)	月平均气温(c)	
月平均日照时数 & 月平均湿度 & 月降雨量(mm)	生长量(cm)	Correlation	1.000	.977
		Significance (2-tailed)	.	.000
		df	0	7
月平均气温(c)	月平均日照时数	Correlation	.977	1.000
		Significance (2-tailed)	.000	.
		df	7	0

根据表10-8可以得出结论：中山柏生长量与气温关系最密切，相关系数0.9774，不相关的概率 $p < 1\%$ ；其次是湿度，相关系数0.7310，假设成立的概率为2.5%；日照时间相关系数0.6318，不相关的概率为6.8%。与降雨量的相关系数是负值，但无统计意义。降雨量过大，会影响其生长。可以看出，偏相关分析结果与简单相关分析结果会有很大区别。

表 10-8 中山柏生长量与四个气候因素的偏相关综合结果

	月均气温	月均湿度	月均日照时数	月降雨量
生长量	.9774	.7310	.6318	-.4906
自由度	( 7 )	( 7 )	( 7 )	( 7 )
不相关概率 p	0.000	0.025	0.068	0.180

### 10.3.4 偏相关分析的过程语句

偏相关分析的过程语句如下：

PARTIAL CORR [VARIABLES=] varlist [WITH varlist] BY varlist [(levels)]

[/varlist...]

[/SIGNIFICANCE={TWOTAIL\*\*}{ONETAIL }]

```
[/STATISTICS=[NONE**] [CORR] [DESCRIPTIVES] [BADCORR] [ALL]]  
[/FORMAT={MATRIX** } {SERIAL } {CONDENSED}]  
[/MISSING=[{LISTWISE**} {ANALYSIS } ] [{EXCLUDE**} {INCLUDE }]]  
[/MATRIX= [IN({* } {file})] [OUT({* } {file})]]
```

## 1. PARTIAL CORR 命令语句

该命令语句调用偏相关分析过程。过程语句由以下几部分组成:

(1) 命令关键字, **PARTIAL CORR** 是定义偏相关分析过程的命令关键字。

(2) 变量定义, 在语句关键字后面必须列出参与相关分析的变量表。必须在 **BY** 后面列出控制变量表。该命令的变量定义部分共有以下几种方式:

① **varlist BY control list** 方式在 “**BY**” 前面是分析变量, 其后为控制变量。分析变量与控制变量不能同名。分析结果给出偏相关矩阵, 矩阵中的每一个相关系数均为在消除了指定的控制变量的影响后, 行变量与列变量的相关系数。

② **varlist WITH varlist BY control list** 方式在 **WITH** 前面是输出结果的行变量, 其后是列变量, **BY** 后面是控制变量。如果 **WITH** 前后的变量数目不同, 则输出结果将是一个矩形矩阵, 该矩阵中的每一个系数均为消除了控制变量影响的行变量与列变量的偏相关系数及其显著性检验结果。

③ **varlist BY control list / varlist** 这种格式的定义在 “/” 后面的变量表相当于格式②中的 **WITH**。该变量表在输出中作为列变量。

④ **VARIABLES=** 在分析变量表 (格式①) 或行变量表 (格式②③) 前面可以加关键字 “**VARIABLE=**”。这是选项, 如果读者对语句比较熟悉, 变量关系清楚, 可以不使用这种形式。

## 2. 子命令

(1) **SIGNIFICANCE** 子命令指定进行单尾 (**ONETAIL**) 还是双尾 (**TWOTAIL**) 检验。使用该子命令, 两种检验必须择其一。不使用此命令, 系统默认进行双尾检验。当相关方向不明时, 应该选择双尾检验; 当相关方向已知时, 应该选择单尾检验。

(2) **STATISTICS** 子命令指定要计算并输出哪些统计量。共有 4 个选项。

① **NONE**, 不要求计算偏相关系数以外的任何统计量。

② **CORR**, 要求计算所有变量 (包括分析变量与控制变量) 间的相关系数、自由度, 同时给出假设检验的概率水平。

③ **DESCRIPTIVES**, 输出各变量 (包括分析变量和控制变量) 的均值与标准差。

④ **ALL** 要求计算并输出①②③各项统计量。指定了 **ALL** 项, 不用指定以上各项。

(3) **FORMAT** 子命令指定输出格式, 共有 3 种选择, 只能选择一种输出方式。

① **MATRIX**, 以矩阵形式输出。对每个变量, 给出它们之间的偏相关系数、自由度和显著性检验的概率。不使用该语句, 则以矩阵方式输出。此为系统默认的输出方式。

② **SERIAL**, 以变量对为单位给出其偏相关系数、自由度和检验的概率水平。

③ **CONDENSED**，以压缩方式输出，只给出偏相关系数。不给出显著性概率和自由度。使用星号表示显著性概率水平。一个星表示概率水平小于等于 5%，两个星表示小于等于 1%。

(4) **MISSING** 子命令指定处理缺失值的方法。

指定了多个分析变量表时，对每个分析表分别处理缺失值。不同的分析表使用不同数目的观测量集。

当成对剔除生效时（关键字 **ANALYSIS**），对特殊的偏相关系数的自由度根据用于计算的观测量最小数确定。

**LISTWISE** 和 **ANALYSIS** 任选一个，每个都可以与 **INCLUDE** 或 **EXCLUDE** 一起使用。系统默认的是 **LISTWISE** 和 **EXCLUDE**。

① **LISTWISE** 按表列剔除带有缺失值的观测量。列在分析表中的任意一个变量带有缺失值的观测量，包括控制变量，都在计算相关系数时剔除。是系统默认的处理方法。计算两个变量的偏相关系数时，这两个变量中任何一个变量带有缺失值，则带有缺失值的观测量从计算中剔除。

② **ANALYSIS** 成对剔除带有缺失值的观测量。在一对变量中的一个或两个变量带有缺失值的观测量都从零阶相关系数的计算中剔除。

③ **EXCLUDE** 剔除所有带有读者定义的缺失值的观测量。

④ **INCLUDE** 带有读者定义的缺失值的观测量参与分析，读者缺失值当作合法值处理。一般不选择此种处理方法。

## 10.4 距离分析

### 10.4.1 距离分析的概念

#### 1. 关于距离分析

距离分析是对观测量之间或变量之间相似或不相似程度的一种测度，是计算一对变量之间或一对观测量之间的广义距离。这些相似性或距离测度可以用于其他分析过程，例如因子分析、聚类分析或多维定标分析，有助于分析复杂的数据集。例如是否可以根据一些特性，如发动机的大小、MPG（每加仑汽油所能行驶的距离）和马力来测度两种汽车的相似性？通过计算汽车间的相似性，可以对这些汽车获得一些认识，哪些汽车彼此类似，哪些汽车彼此不同。更正规的分析，可以考虑对相似性使用分层聚类或多元定标分析去探测深层结构。

#### 2. 有关的统计量

##### (1) 不相似性测度

① 对等间隔数据的不相似性（距离）测度可以使用的统计量有：**Euclidean distance**



欧几里得 (欧氏) 距离、Squared Euclidean Distance 欧氏距离平方、Chebychev 切比雪夫、Block 区组、Minkowski 明可斯基或 Customized 自定义统计量。

② 对计数数据, 使用卡方或斐方 ( $\Phi^2$ )。

③ 对二值变量 (只有两种取值) 数据, 使用欧氏距离、欧氏距离平方、尺寸差异、模式差异、方差、形或兰斯和威廉斯等距离统计量。

这些统计量的计算方法参见 13.3.3 小节关于聚类方法选项 (Method) 相关内容。

(2) 相似性测度

① 等间隔数据使用统计量皮尔逊相关或余弦。

② 测度二元数据相似性使用的统计量有二十余种。算法参见附录 A。

在 SPSS 中的距离分析属于专业统计分析过程 (Professional Statistics Options), 是选项。如果没有安装, 则在菜单中不会有调用该过程的菜单项。

距离分析分为观测量之间距离的分析和变量之间距离的分析。

## 10.4.2 距离分析过程

SPSS 的距离分析过程提供相似性和不相似性两种分析方法。

1. 按 Analyze→Correlate→Distance 顺序单击菜单项, 展开了 Distances 距离分析的主对话框, 如图 10-8 所示。

在主对话框中可以看到 Compute Distances 组中系统默认 Between cases, 在 Measure 组中系统默认 Dissimilarities, 即观测量间的不相似性测度。在 Measure 按钮旁边显示的 Euclidean distance 表明使用欧几里得 (欧氏) 距离测度观测量间的不相似性。

2. 指定分析变量和标识变量

对于观测量间的距离分析至少指定一个分析变量和一个标识变量。在源变量栏中选择分析变量, 将其移至 Variables 矩形框中, 选择一个标识变量, 将其移至下面一个 Label Cases by: 矩形框中, 如图 10-8 所示。

3. 主对话框中的选项

(1) Compute Distances 计算距离栏

① Between cases, 计算每对观测量间的距离。

② Between variables, 计算每对变量间的距离。

(2) 在 Measure 栏中选择测度距离的类型与方法。

① Dissimilarities, 计算不相似性矩阵, 此为系统默认的类型。系统默认使用欧氏距离测度其不相似性。

② Similarities, 计算相似性矩阵。系统默认使用 Pearson 相关进行相似性测度。

在 Measure 栏中选择了一种测度类型后, 系统默认的计算方法显示在 Measure 按钮右侧。可以单击 Measure 按钮打开相应的对话框, 进一步选择计算方法或统计量。返回主对话框后, 被选中的计算方法显示在按钮旁边。

单击 Measure 按钮打开如图 10-9 所示对话框，指定不相似性测度的计算方法选项。

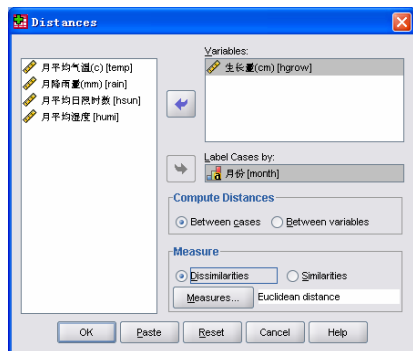


图 10-8 距离分析的主对话框

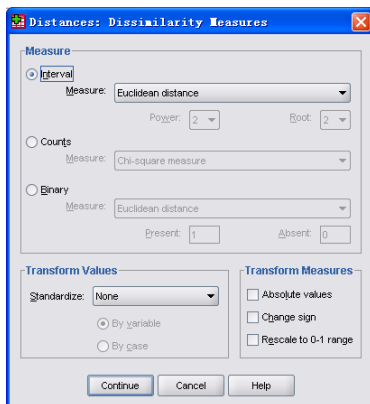


图 10-9 不相似距离测度选项对话框

#### 4. Dissimilarity Measure 有关不相似性测度的选项

##### (1) 不相似性测度方法选项 Measure 栏

选择一种测度需要首先选择数据类型，然后，在选中的数据类型组的下拉菜单中选择与数据类型一致的可用的测度方法，对话框见图 10-9。数据类型及其可以使用的测度如下：

① **Interval**，等间隔变量（即指连续变量）选项，各选项的详细说明见 11.3.2 小节中的有关内容。

② **Counts**，计数变量选项。在选择了该选项后，可以在展开的下拉表中选择不相似性测度。各选项的详细说明见 14.3.2 小节中的有关内容。

③ **Binary**，二值变量（表示某种特性有、无的变量）选项。选择该项后激活其下的其他选项。在 **Present** 框中输入表明特性存在的变量值，在 **Absent** 框中输入表明不存在某特性的变量值。系统默认的变量值是用 1 表明特性存在，用 0 表明特性不存在。对于二值变量的各选项的详细说明见附录 A 中的有关内容。

##### (2) Transform Values 转换数值栏

该栏允许在进行近似计算之前对观测量或变量进行标准化，但对二元变量不能进行标准化。

① **Standardized** 框中选择标准化的方法，各选项的详细说明见附录。

② 以上除了 **None** 选项外，选择其他任意一种标准化的方法，均应同时指定标准化对象，共有两个选项：

- **By variable**，即对变量进行标准化。
- **By cases**，即对观测量进行标准化。

(3) Transform Measures 转换测度栏

转换测度组选择在距离测度计算完成后，对距离测度的结果进行转换的方法。共有 3 种方法可以选择。3 种转换方法可以同时选择。

① Absolute values，即对距离取绝对值。当符号表明的是相关的方向，且仅对相关的数值感兴趣时使用这种转换。

② Change sign，改变符号。把相似性测度值转换成不相似性测度值或相反。使用这种转换，通过加负号，颠倒距离测度的顺序。

③ Rescale to 0-1 range，即先减去最小值，然后除以范围（最大值减最小值）使距离标准化。对已经按有意义的方法标准化的测度，一般不再使用此方法进行转换。

5. Similarity Measure 相似性测度的选项

在主对话框中选择相似性测度，并使用鼠标单击 Measure 按钮，展开 Distances: Similarity Measure 相似性测度对话框，如图 10-10 所示。

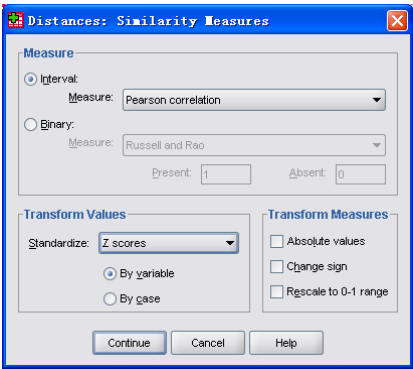


图 10-10 相似性测度选项对话框

(1) Measure 相似性测度方法的选项栏

有关相似性测度方法的选项与不相似性测度一样，在选择具体的测度方法之前必须首先选择变量类型。然后，在选中的实际类型组的下拉菜单中选择与实际类型一致的可用的测度。进行相似性测度的数据类型只有两种：等间隔变量和二元变量。与这两个类型相应的可以选择的测度方法如下：

① Interval，等间隔变量选项，各选项的详细说明见附录 A。

② Binary，对二元数据的相似性测度。SPSS 为每对项目构造一个 2×2 的列联表。可用的测度落入以下 4 类中：匹配系数、条件概率、可预测性测度和其他测度。可以从下拉表中选择一种测度。在进行测度方法选择之前应该指定表明某特点出现 Present 和不出现 Absent 的变量值，系统默认，特点出现其值为 1，特点不出现其值为 0。读者可以指定其他整数表明特性的出现与不出现，SPSS 将忽略其他值。各选项的详细说明见附录 A。

(2) Transform Values 栏，转换数值，参见本小节 4.(2)。

(3) Transform Measures 栏，转换测度，参见本小节 4.(3)。

10.4.3 距离分析实例

【例 6】观测量间的相似性分析例题。

(1) 这里仍使用 data10-04，四川绵阳地区中山柏生长的数据 data10-03。分析不同月份间生长量之间的距离以便分析各月份生长量间的相似或不相似性。读入数据文件，按下述步骤操作。

(2) 按 Analyze→Correlate→Distance 顺序单击菜单项, 展开 Distance 距离分析的主对话框, 如图 10-7 所示。

(3) 指定分析变量和标识变量。选择月生长量 hgrow 作为分析变量, 将其移至 Variables 矩形框中。选择月份 Month 作为标识变量, 将其移至下面一个 Label Cases by: 矩形框中。

(4) 按 OK 按钮, 提交运行。执行的程序与结果如下:

PROXIMITIES

Hgrow

/ID=month

/VIEW=CASE

/MEASURE= EUCLID

/STANDARDIZE= NONE.

①

②

③

④

⑤

⑥

① PROXIMITIES 是调用距离分析过程的命令。

② Hgrow 是指定的分析变量。

③ ID 子命令指定标识变量是 month。

④ VIEW 子命令指定计算观测量间的欧氏距离, 表明观测量间的不相似性。

⑤ MEASURE 子命令指定计算观测量间的欧氏距离, 表明观测量间的不相似性。

⑥ STANDARDIES 子命令指定不进行变量的标准化。

(5) 输出结果见表 10-9 和表 10-10。

表 10-9 观测量统计处理简明表

Case Processing Summary					
		Cases			
		Valid		Missing	
		N	Percent	N	Percent
		12	100.0%	0	.0%

表 10-10 观测量间的欧氏距离

Proximity Matrix												
	Euclidean Distance											
	1: 1	2: 2	3: 3	4: 4	5: 5	6: 6	7: 7	8: 8	9: 9	10:10	11:11	12:12
1: 1	.000	.490	1.490	10.790	12.990	16.290	17.990	19.290	14.790	10.290	7.990	.990
2: 2	.490	.000	1.000	10.300	12.500	15.800	17.500	18.800	14.300	9.800	7.500	.500
3: 3	1.490	1.000	.000	9.300	11.500	14.800	16.500	17.800	13.300	8.800	6.500	.500
4: 4	10.790	10.300	9.300	.000	2.200	5.500	7.200	8.500	4.000	.500	2.800	9.800
5: 5	12.990	12.500	11.500	2.200	.000	3.300	5.000	6.300	1.800	2.700	5.000	12.000
6: 6	16.290	15.800	14.800	5.500	3.300	.000	1.700	3.000	1.500	6.000	8.300	15.300
7: 7	17.990	17.500	16.500	7.200	5.000	1.700	.000	1.300	3.200	7.700	10.000	17.000
8: 8	19.290	18.800	17.800	8.500	6.300	3.000	1.300	.000	4.500	9.000	11.300	18.300
9: 9	14.790	14.300	13.300	4.000	1.800	1.500	3.200	4.500	.000	4.500	6.800	13.800
10:10	10.290	9.800	8.800	.500	2.700	6.000	7.700	9.000	4.500	.000	2.300	9.300
11:11	7.990	7.500	6.500	2.800	5.000	8.300	10.000	11.300	6.800	2.300	.000	7.000
12:12	.990	.500	.500	9.800	12.000	15.300	17.000	18.300	13.800	9.300	7.000	.000

This is a dissimilarity matrix

表 10-9 是对观测量有效值和缺失值进行的统计。

表 10-10 以矩阵形式给出了两两观测量间变量 **hgrow** 的欧氏距离, 即每两个月份间的中山柏生长量间的差值, 这是不相似矩阵, 行列之间数值越大的不相似性越强。显然, 1 月与 8 月生长量最不相似, 其欧氏距离值为 19.290。1 月、2 月生长量不相似性最小, 值为 0.490。12 月、2 月、3 月生长量的不相似性和 4 月、10 月的生长量不相似性仅次于 1 月、2 月, 值为 0.5。

在进行观测量间不相似性分析时, 可以指定若干个分析变量, 即根据指定变量组分析观测量间的不相似性。标识变量只能指定一个。

**【例 7】变量间的不相似性例题。**

对于连续变量间的相似性计算, 往往使用 **Pearson** 相关, 这与两个变量间的简单相关分析没有区别。本例仍使用数据 **data010-03**, 比较相似性与不相似性的结果。

(1) 按 **Analyze→Correlate→Distance** 顺序单击菜单项, 展开 **Distances** 的主对话框。

(2) 指定分析变量: 月平均的: 气温 **temp**、降雨量 **rain**、日照时间 **hsun**、湿度 **humi**。选择它们并将其移至 **Variables** 矩形框中。

(3) 在 **Compute Distances** 栏中选择 **Between variables**, 在 **Measure** 栏中选择 **Dissimilarities**, 要求进行变量间的不相似性分析。

(4) 单击 **Measure** 按钮。

① 在 **Measure** 对话框中选择 **Interval**, 因为所选择的变量均为等间隔测度的变量。在下拉列表中选择 **Euclidean distance**, 因为只有在相似性分析的菜单中才有相关分析的选项, 而不相似性分析不计算相关矩阵, 只计算欧氏距离或其他距离。

② 因为所选择的分析变量测度的单位不同, 因此要对变量进行标准化。在 **Transform Values** 栏选择 **By variable**, 在下拉列表中选择 **Z scores**, 对变量进行均值为 0, 标准差为 1 的标准化。

(5) 运行的程序语句如下。

```
PROXIMITIES                                ①
temp rain hsun humi
/VIEW=VARIABLE                              ②
/MEASURE= EUCLID                            ③
/STANDARDIZE= VARIABLE Z .                  ④
```

语句①调用进行距离分析的过程, 指定分析变量; 语句②、③ 要求计算变量间的欧氏距离。④要求在计算距离之前对变量进行标准化, 各变量标准化到 Z 分数。运行结果见表 10-11。

表 10-12 是选择对变量使用 **Pearson Correlation** 进行相似性测度的结果。即上面程序的③ 改为 **/MEASURE=CORRELATION** 运行结果。

比较两种分析结果, 可以看出结果是一致的。

表10-11不相似性分析的结果，是欧氏距离矩阵；表10-12是相似性分析的Pearson相关矩阵；相似性越强，相关系数越大，不相似性距离越小。例如月平均气温与月降雨量相关系数最大，为0.715；在不相似性的距离矩阵中，这两个变量间的距离最小，为2.505。相反，在相似性测度的相关矩阵中，相关系数最小的是月平均湿度与月平均日照时间，为-0.051，它们的不相似性测度中的欧氏距离却是最大的，值为4.808。

表 10-11 变量间的不相似性测度标准化后的欧氏距离

Proximity Matrix				
	Euclidean Distance			
	月平均气温(c)	月降雨量(mm)	月平均日照时数	月平均湿度
月平均气温(c)		2.505	2.609	3.947
月降雨量(mm)	2.505		2.561	3.680
月平均日照时数	2.609	2.561		4.808
月平均湿度	3.947	3.680	4.808	

This is a dissimilarity matrix

表 10-12 变量间的相似性测度，相关分析结果

Proximity Matrix				
	Correlation between Vectors of Values			
	月平均气温(c)	月降雨量(mm)	月平均日照时数	月平均湿度
月平均气温(c)		.715	.690	.292
月降雨量(mm)	.715		.702	.384
月平均日照时数	.690	.702		-.051
月平均湿度	.292	.384	-.051	

This is a similarity matrix

关于距离分析的过程语句参见 13.4.6 小节中的 PROXIMITIES 过程语句的有关内容。

## 习 题 10

1. 什么是两个变量间的线性相关？两个变量间的相关系数的数值范围是什么？负相关系数反映的是两个变量数值间的什么样的关系？
2. SPSS提供了几个求相关系数的方法？各适合分析什么样的变量？
3. 在data10-05中记录了29个被试者的身高、体重、肺活量的数据，试分析肺活量与哪个因素线性相关程度更高。说明为什么要计算偏相关？
4. 在data10-02中是474名职工的职务等级jobcat、起始工资salbegin、现工资salary、受教育程度educ、本单位工作经历（月）jobtime、以前工作经历（月）prevexp，id为职工编号。分析该公司起始工资的确定与什么因素有关。当前工资与什么因素有关。
5. data10-06是某公司太阳镜销售情况。分析销售量与平均价格、广告费用和日照时间之间的关系。作图协助分析。此题使用偏相关分析是否有实际意义？

# 第 11 章 回 归 分 析

回归分析是在自然科学、社会科学等领域中具有广泛应用的统计方法。变量与变量之间的关系分为确定性关系和非确定性关系两类。函数表达确定性关系。研究变量间的非确定关系，构造变量间经验公式的数理统计方法称为回归分析。

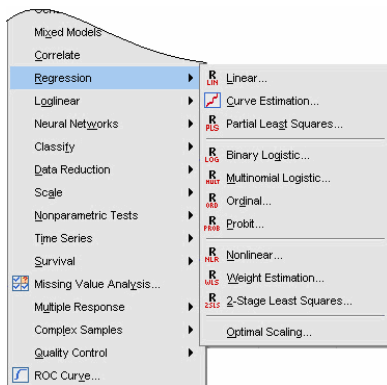


图 11-1 回归菜单

在 Analyze 的 Regression 菜单中，回归分析过程有以下几种，如图 11-1 所示。

Linear 线性回归、Curve Estimation 曲线估计、Partial Least Squares Regression 偏最小二乘回归、Binary Logistic 二分变量逻辑斯谛回归、Multinomial Logistic 多分变量逻辑斯谛回归、Ordinal 定序回归、Probit 概率单位回归、Nonlinear 非线性回归、Weight Estimation 加权估计、2-Stage Least Squares 两段最小二乘法、Optimal Scaling 最优编码尺度回归。

## 11.1 线 性 回 归

自变量与因变量之间呈线形关系时，我们可以构造线性回归方程。根据参与线性回归的自变量个数的多少，可将线性回归分为一元线性回归和多元线性回归。

### 11.1.1 一元线性回归

#### 1. 一元线性回归方程

只有 1 个自变量的线性回归，称一元线性回归，又称直线回归。其分析的任务就是根据若干个观测 $(x_i, y_i)$   $i=1, 2, \dots, n$  找出描述两个变量  $x$  与  $y$  之间关系的直线回归方程  $\hat{y} = a + bx + \varepsilon_i$ 。其中  $\hat{y}$  是实测变量  $y$  的估计值。求最优线性回归方程  $\hat{y} = a + bx$ ，常用的方法是最小二乘法，也就是使该直线与各点的纵向垂直距离最小，即实测值  $y$  与预测值  $\hat{y}$  之差的平方和  $\sum (y - \hat{y})^2$  达到最小。 $\sum (y - \hat{y})^2$  也称为剩余（残差）平方和。因此求回归方程  $\hat{y} = a + bx$  的问题，归根结底就是求  $\sum (y - \hat{y})^2$  取得最小值时  $a$  和  $b$  的问题。 $a$  称为截距， $b$  为回归直线的斜率，它们又称回归系数。

#### 2. 一元线性回归方程的假设理论

德国数学家高斯提出 5 个假设理论，满足这些假设的线性模型称为古典线性模型。

(1) 正态性假设: 随机误差项  $\varepsilon_i$  服从均值为 0, 方差为  $\sigma^2$  的正态分布。

(2) 等方差假设: 对所有  $x_i$ ,  $\varepsilon_i$  的条件方差同为  $\sigma^2$ , 且  $\sigma$  为常数, 即

$$\text{Var}(\varepsilon_i/x_i) = \sigma^2$$

(3) 独立性假设即零均值假设: 在给定  $x_i$  的条件下,  $\varepsilon_i$  的条件期望值为 0, 即

$$E(\varepsilon_i) = 0$$

(4) 无自相关性假设: 随机误差项  $\varepsilon$  的逐次观察值互不相关, 即

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j)$$

(5)  $\varepsilon$  与  $x$  的不相关性。假设随机误差项  $\varepsilon_i$  与相应的自变量  $x_i$  对因变量  $y$  的影响相互独立。换言之, 两者对因变量  $y$  的影响是可以区分的。即  $\text{Cov}(\varepsilon_i, x_i) = 0$ 。

### 3. 一元线性回归方程的检验

根据原始数据, 求出回归方程后就需要对回归方程进行检验。检验的假设是总体回归系数为 0。可以选用下述前三种方法中的任意一种。另外, 还要检验回归方程的预测效果如何。

#### (1) 回归系数的显著性检验

① 对斜率检验的假设是, 总体回归系数  $b=0$ 。检验该假设的  $t$  值计算公式是

$$t = \frac{b}{SE_b}$$

② 对截距检验的假设是, 总体回归方程截距  $a=0$ 。检验该假设的  $t$  值计算公式是

$$t = \frac{a}{SE_a}$$

在两公式中,  $SE_b$  是回归系数的标准误。 $SE_a$  是截距的标准误。

(2)  $R^2$  判定系数。它是判定线性回归直线拟合优度的重要指标。公式为

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

表明判定系数等于回归平方和在总平方和中所占的比率, 体现了回归模型所解释的因变量变异的百分比。如果  $R^2=0.775$ , 说明变量  $y$  的变异中有 77.5 % 是由变量  $x$  引起的。 $R^2=1$ , 表明因变量与自变量为函数关系。 $R^2=0$ , 表示自变量与因变量无线性关系。

#### (3) 方差分析

因变量观测值与均值之间差异的偏差平方和  $SS_t$  由两个部分组成, 表示为  $SS_t = SS_r + SS_e$ 。其中回归平方和  $SS_r$ , 反映了自变量  $x$  的重要程度; 残差平方和  $SS_e$ , 反映了实验误差以及其他意外因素对实验结果的影响。这两部分除以各自的自由度, 得到它们的均方

$$F = \frac{\text{回归均方}}{\text{残差均方}} = \frac{\sum (\hat{y} - \bar{y})^2 / p}{\sum (y - \hat{y})^2 / (n - p - 1)}$$



当  $F$  值太大时, 拒绝  $b=0$  的假设。

#### (4) Durbin-Watson 检验

在对回归模型的诊断中, 需要诊断回归模型中误差项的独立性。如果误差项不独立, 那么对回归模型的任何估计与假设所做出的结论都是不可靠的。

其参数称为  $DW$  或  $D$ 。 $D$  的取值范围是  $0 < D < 4$ , 统计学意义如下:

- ① 当残差与自变量互为独立时,  $D \approx 2$ 。
- ② 当相邻两点的残差为正相关时,  $D < 2$ 。
- ③ 当相邻两点的残差为负相关时,  $D > 2$ 。

(5) 残差图示法。在直角坐标系中, 常以预测值  $\hat{y}$  为横轴, 以  $y$  与  $\hat{y}$  之间的误差  $e_i$  (或学生式残差值) 为纵轴, 绘制残差的散点图。如果散点呈现出明显的规律性, 则认为存在自相关性, 或者存在非线性、非常数方差的问题, 如图 11-2(a)~(d)所示。

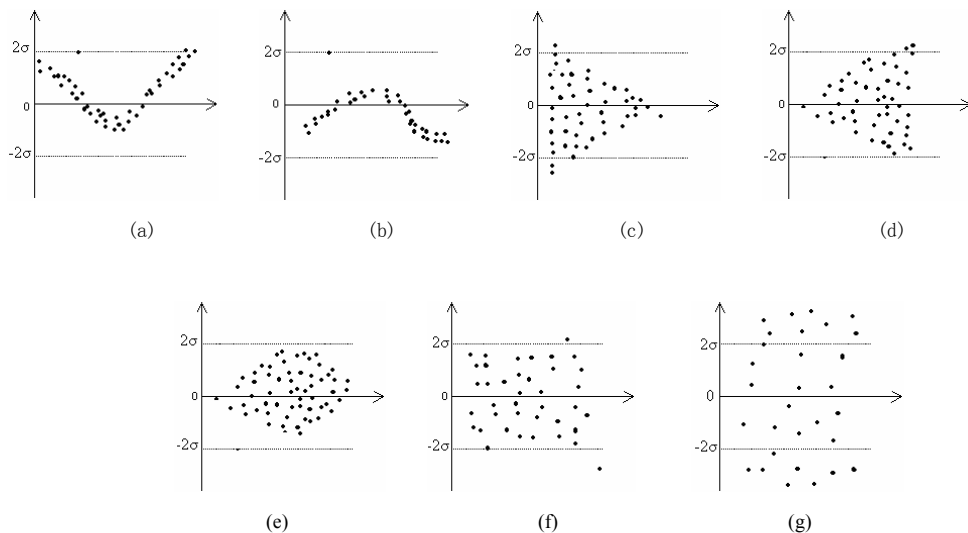


图 11-2 各种残差与预测值关系示意图

如果残差与因变量的关系类似图 11-2(a)或(b), 则需要对因变量或自变量进行变换。如果散点呈随机分布, 则认为残差与因变量之间相互独立, 如图 11-2(f)所示。

利用残差图还可以判断模型拟合效果。在图中, 如果各点呈随机状, 并绝大部分落在  $\pm 2\sigma$  范围 (68% 的点落在  $\pm \sigma$  之中, 96% 的点落在  $\pm 2\sigma$  之中) 内, 说明拟合效果较好, 如图 11-2(f)所示。如果大部分点落在  $\pm 2\sigma$  范围之外, 说明拟合效果不好, 见图 11-2(g)。

### 11.1.2 多元线性回归

#### 1. 多元线性回归的概念

根据多个自变量的最优组合建立回归方程来预测因变量的回归分析称为多元回归分析。多元线性回归分析拟合后的方程为  $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \Lambda + b_nx_n$ 。

其中  $\hat{y}$  为根据所有自变量  $x$  计算出的估计值,  $b_0$  为常数项,  $b_1, b_2, \dots, b_n$  称为  $y$  对应于  $x_1, x_2, \dots, x_n$  的偏回归系数。偏回归系数表示假设在其他所有自变量不变的情况下, 某一个自变量变化引起因变量变化的比率。

多元线性回归模型也必须满足 9.1.1 节中所述的假设理论。

## 2. 多元线性回归分析中的统计指标

(1) 复相关系数  $R$  表示自变量  $x_i$  与因变量  $y$  之间线性关系密切程度的指标, 取值范围在 0~1 之间。其值越接近 1, 表示线性关系越强; 越接近 0, 表示线性关系越差。

### (2) $R^2$ 判定系数与校正 $R^2$ 判定系数

在多元回归中也使用  $R^2$  判定系数解释回归模型中自变量的变异在因变量变异中所占的比率。但是, 在多元回归中判定系数的值会随着进入回归方程的自变量的个数  $n$  或样本容量的大小的增加而增大。为了消除自变量的个数以及样本量的大小对判定系数的影响, 引进了校正  $R^2$  (Adjusted R Square)。校正  $R^2$  判定系数的公式是

$$\text{Adjusted } R^2 = 1 - \frac{\sum (y - \hat{y})^2 / (n - k - 1)}{\sum (y - \bar{y})^2 / (n - 1)}$$

其中  $k$  为自变量的个数,  $n$  为观测量数目。可以看出, 自变量数大于 1 时, 其值小于  $R^2$  判定系数。自变量数越多, 与  $R^2$  判定系数的差值越大。

### (3) 零阶相关系数、部分相关系数与偏相关系数

① 零阶相关系数 (Zero-Order) 表示: 各自变量与因变量之间的简单相关系数。

② 部分相关 (Part Correlation) 表示: 在排除了其他自变量对  $x_i$  的影响后, 当一个自变量进入回归方程模型后, 复相关系数的平方的增加量。

③ 偏相关系数 (Partial Correlation) 表示: 在排除了其他变量的影响后, 自变量  $x_i$  与因变量  $y$  之间的相关程度。部分相关系数小于偏相关系数。偏相关系数也可以用来作为筛选自变量的指标, 即通过比较偏相关系数的大小, 判别哪些变量对因变量具有较大的影响力。

## 3. 多元线性回归分析的检验

可以利用残差分析, 检验建立的回归模型是否很好地拟合了原始数据。还可以对回归方程中各自变量的系数进行检验, 以便在回归方程中保留那些有效影响因变量  $y$  值的自变量。

(1) 方差分析是对整个回归方程的显著性检验。检验的假设是: 总体的回归系数均为 0。使用统计量  $F$  进行检验, 其原理与一元回归的方程分析原理相同。

(2) 偏回归系数与常数项的检验, 检验的假设是: 总体中回归方程各自变量偏回归系数为 0, 常数项为 0。检验使用  $t$  统计量。偏回归系数和常数项的  $T$  检验的公式分别是

$$t = \frac{\text{偏回归系数}}{\text{偏回归系数的标准误}}, \quad t = \frac{\text{常数项}}{\text{常数项的标准误}}$$

(3) 方差齐性检验, 检验方差齐性是指残差的分布是常数, 与自变量或因变量无关。一般用绘制因变量预测值与学生式残差的散点图来检验。残差应随机地分布在一条穿过零点的水平直线的两侧。

(4) 残差的正态性检验。希望残差完全服从于正态分布也是不现实的, 即使存在很理想的总体数据, 其样本的残差的分布也只能是近似于正态分布。

最直观、最简单的方法是作残差的直方图和累积概率图。

累积概率图 (P-P 图) 是用来判断一个变量的分布是否与一个指定的分布一致。如果两种分布基本相同, 那么在 P-P 图中的点应该围绕在一条斜线的周围。通过观察残差 (曲线) 在假设直线 (正态分布) 周围的分布, 可以判断是否符合正态分布。

### 11.1.3 异常值、影响点、共线性诊断

#### 1. 异常值的查找

异常值是指标准化残差过大的观测量, 在 SPSS 软件中, 默认的判据是标准化残差的绝对值大于 3。

#### 2. 影响点的查找

因为影响点对参数估计的结果有较大的影响, 所以要仔细地考虑在模型拟合时是保留还是剔除影响点。要注意, 影响点的非标准化残差并不太大, 因此需要仔细研判。

识别影响点的有效方法, 是比较一个观测量存在于回归方程时与不存在于回归方程式时残差的变化。主要的指标有: 标准化残差、非标准化残差、学生式剔除残差、学生式残差、剔除残差、Mahalanobis 距离、中心点杠杆值、Cook 距离、协方差比。

需要使用几个指标综合判断某一观测量是否为影响点, 使用个别指标可能会错判。

##### (1) 判别影响点的指标

① **Dresid**, 剔除残差。排除一个被认为是影响点的观测量, 回归分析的残差值。

② **Sdresid**, 学生化残差。残差除以它的标准误, 其值大于 2 时, 应予以重视。

③ **Cook 距离**。它是对当一个被认为是影响点的观测量被删除后, 其他所有观测量残差的变化量的测度。此值越大, 表示这个被认为是影响点的观测量的影响力越大。

④ **Mahalanobis 距离**。测定某一自变量观测量与同一自变量所有观测量平均值差异的统计量。此值越大说明该观测量为影响点的可能性越大。

⑤ **Leverage values**, 中心点杠杆值。当回归方程含有一个以上的自变量时, 用来检测影响点的标准。其值在  $0 \sim (N-1)/N$  之间变化, 杠杆值为 0 时, 说明此观测值对回归方程没有影响; 杠杆值接近  $(N-1)/N$ , 说明此观测量对回归方程的贡献很大。从理论上说, 希望数据所有的观测量的杠杆值都接近于中心点杠杆平均值  $P/N$  ( $P$  为自变量数目), 当杠杆值大于  $2 \times P/N$  时, 说明此观测量的影响力很大。

⑥ **Covariance Ratio**, 协方差比, 它用来衡量某个观测量是否对回归系数有显著的影响。

响。当协方差比的值接近 1 时,表明此点的观测量不是影响点。

国外有学者建议,当  $| \text{协方差比} - 1 | \geq 3P/N$  时,这个观测量可以被视为影响点。

### (2) 利用回归系数的变化检验影响点

Belsley 给出的建议是:仔细检查某一观测量在与不在模型中前后变化的标准化  $\beta$  值,如果大于  $2/\sqrt{N}$  ( $N$  为观测量的数目),那么此观测量就有可能是影响点。

### (3) 利用预测值来检测影响点

如果从模型中删除某一个观测量后,其标准化预测值大于  $2/\sqrt{P/N}$  ( $P$  为自变量的个数,  $N$  为观测量数) 时,此观测量有可能是影响点。

## 3. 共线性问题

在回归方程中,各自变量对因变量虽然都是有意义的,但某些自变量彼此相关,即存在共线性的问题。这给评价自变量的贡献率带来困难。因此,需要对回归方程中的变量进行共线性诊断,并且确定它们对参数估计的影响。

共线性分为精确共线性与近似共线性。如果存在一些常数  $c_0$ 、 $c_1$ 、 $c_2$ ,使得等式  $c_1x_1+c_2x_2=c_0$ ,对数据中所有的观测量都成立,则两个自变量  $x_1$  与  $x_2$  之间的关系为精确共线性;如果这个等式近似成立,那么两个自变量  $x_1$  与  $x_2$  之间的关系为近似共线性。

在只有两个自变量的情况下,  $x_1$  与  $x_2$  共线性体现在两自变量间相关系数  $r_{12}$  上。精确共线性时  $r_{12}^2=1$ ,当它们之间不存在共线性时,  $r_{12}^2=0$ 。 $r_{12}^2$  越接近于 1,共线性越强。

当自变量多于两时,  $x_i$  与其他自变量  $x$  之间的复相关系数的平方体现共线性,称它为  $R_i^2$ 。它的值越接近 1,说明自变量之间的共线性程度越大。

当一组自变量精确共线性时,必须删除引起共线性的一个和多个自变量,否则不存在系数唯一的最小二乘估计。因为删除的自变量并不包含任何多余的信息,所以得出的回归方程并没有失去什么。当为近似共线性时,一般是将引起共线性的自变量删除,但需要掌握的原则是:务必使丢失的信息最少。识别共线性的统计量有以下几个:

① 容许度 (Tolerance) 定义为  $\text{Tol}_i=1-R_i^2$ ,其值介于 0~1 之间。其值越小,自变量  $x_i$  与其他自变量  $x$  之间的共线性越强。使用容许度作为共线性量度指标的条件比较严格,观测量一定要近似于正态分布。

② 方差膨胀因子 (VIF),方差膨胀因子 (VIF) 定义为  $\text{VIF}_i=1/(1-R_i^2)$ ,是容许度的倒数,其值介于 1~ $\infty$  之间。其值越大,自变量之间存在共线性的可能性越大。

有专家认为,容许度小于 0.1 或 0.2 或者 VIF 值大于 5 或 10 可以认为存在共线性问题。读者可以参考。

③ 特征值 (Eigenvalues),当若干特征值较小并且接近 0 时,说明某些变量之间存在很高的相关性。这些变量的观测量出现较小的变化时,都会导致回归系数较大的变化。

④ 条件指数 (Condition Index) 是在计算特征值时产生的一个统计量,其值越大,说明自变量间的共线性的可能性越大。一般认为,条件指数  $\geq 15$  时可能存在共线性问题,当条件指数  $\geq 30$  时存在严重的共线性问题。 $\text{Condition Index}=\sqrt{\text{最大特征值}/\text{第}i\text{个特征值}}$ 。

⑤ 方差比例（Variance Proportions），同一序号的特征值对应的变量的方差比例。比例越大，其共线性的可能性越大。

⑥ 常用的共线性问题的解决方法：

- 从产生共线性问题的自变量中剔除不重要的自变量。
- 增加样本量。
- 重新抽取样本数据。不同样本的观测量的共线性是不一致的，所以重新抽取样本数据有可能减少共线性问题的严重程度。

11.1.4 变非线性关系为线性关系

因变量与自变量的关系不是线性关系，但利用其他方法也未能很好地拟合数据时，就需要进行数据的非线性到线性关系的转换。如果因变量或残差不符合假设条件，也需要进行转换。非线性转换为线性关系的原则及方法的统计学知识已经超出本书范围，读者可以参考有关书籍，在此仅给予提示。

1. 当残差的分布呈现正偏态时，对因变量进行对数转换。当残差的分布呈现负偏态分布时，采用平方根转换。
2. 如果残差的方差呈现不稳定状态，可用表 11-1 的方法进行校正，注意适用条件。

表 11-1 变量转换公式表

转换方法	使用条件	注 释
$\sqrt{y}$	$\text{Var}(e_i) \propto E(y_i)$	因变量服从泊松分布
$\sqrt{y + \sqrt{y + 1}}$	$\text{Var}(e_i) \propto E(y_i)$	某些因变量的值为 0，或者很小
$\text{Log } y$	$\text{Var}(e_i) \propto [E(y_i)]^2, y > 0$	因变量的值的范围很大
$\text{Log } (y + 1)$	$\text{Var}(e_i) \propto [E(y_i)]^2$	因变量的某些值为 0
$1/y$	$\text{Var}(e_i) \propto [E(y_i)]^4$	因变量的值集中在 0 的附近，当自变量明显降低时，因变量出现较大的值。 例如：自变量是治疗某病的药剂量，因变量是反应时间。
$1/(y + 1)$	$\text{Var}(e_i) \propto [E(y_i)]^4$	某些自变量为 0 的情况
$\arcsin \sqrt{y}$	$\text{Var}(e_i) \propto E(y_i)(1 - (y_i))$	用于二项比例 ( $0 \leq \text{因变量} \leq 1$ )

注：符号  $\text{var}(e_i)$  为  $e_i$  的方差， $e_i$  为第  $i$  个观测量的统计误差。 $E(y_i)$  为随机变量  $y_i$  的算术平均数。

当方差随着因变量的增大或减小而变化时， $\sqrt{y}$ 、 $\log(y)$  与  $1/y$  都是可以选用的方法，但是它们转换的力度却是依次递增的；当因变量是直到某一事件发生和完成的时间，则常使用倒数或逆变换；当因变量的数据中出现 0 或负数时，为了避免对数或开根号没有意义的情况出现，常采用  $(y + \text{常数})$  的方法，常数一般取 1；在经济学研究方面的  $\log(y)$  是一种常用的方法。

3. 非线性关系转变为线性关系的方法

非线性数据转变为线性数据的方法主要包括：取对数、倒数和平方根。注意，并非

所有的函数都是可以线性化的。

(1) 当回归方程有可能是多项式方程, 如  $y=x^2+3x+1$  时, 可以取平方根或倒数。

(2) 当要建立的回归方程未知时, 可以利用散点图中发现规律, 进行转换, 见表 11-2。

### 11.1.5 线性回归过程

#### 1. 数据要求

(1) 自变量与因变量应该是数值型变量, 类似研究领域、居住地区、信仰等分类变量应重新编码为哑变量或者其他类型的对比变量。

(2) 假设。对自变量的每一个值, 因变量的分布必须是正态的。因变量方差的分布对所有自变量的值都应该是一个常数。因变量和每个自变量之间的关系应该是线性的, 所有观测应该是独立的。

在进行回归分析之前, 最好用图形探索因变量随自变量变化的趋势, 以便确定数据是否适合线性模型。通过散点图还可以发现异常值。

#### 2. 建立线性模型的操作步骤

(1) 按 **Analyze**→**Regression**→**Linear** 顺序打开如图 11-3 所示的主对话框。

(2) 在源变量框中选择一个因变量进入 **Dependent** 框, 选择一个或多个自变量进入 **Independent(s)**框。

可以利用 **Previous** 与 **Next** 按钮切换, 选择不同的自变量组构建不同的模型; 每个模型中可以对不同自变量组采用不同的分析方法, 如有的自变量组采用 **Enter**, 有的采用 **Stepwise**。构建的模型按顺序保存第  $n$  个模型中。

(3) 在 **Method** 框中选择回归分析方法。

① **Enter**, 强行进入法。即所选择的自变量全部进入回归模型。这是默认方式。

② **Stepwise**, 逐步回归法。根据在 **Options** 对话框中所设定的判据, 选择符合判据的自变量且对因变量贡献最大的进入回归方程。然后将模型中符合剔除判据的变量移出模型, 重复进行直到回归方程中的自变量均符合进入模型的判据, 模型外的自变量都不符合进入模型的判据为止。

③ **Remove**, 消去法。先建立全模型, 再进行剔除, 一步就剔除部分自变量。

④ **Forward**, 向前选择法。从模型中无自变量开始, 根据在 **Options** 对话框中所设定的判据, 每次将一个最符合判据的变量引入模型, 直至所有符合判据的变量都进入模

表 11-2 转换为线性的常用方法

变化方法		回 归 式
logy	logx	$y=\alpha\chi$
logy	x	$y=ae^{bx}$
y	logx	$y=\alpha+\beta\log x$
1/y	1/x	$y=x/(a+\beta)$
1/y	x	$y=1/(a+\beta x)$
y	1/x	$y=\alpha+\beta(1/x)$

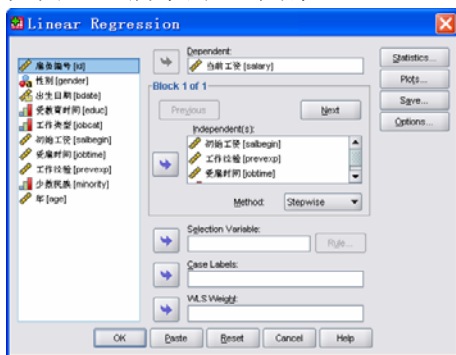


图 11-3 线性回归主对话框

型为止。第一个引入回归模型的变量应该是与因变量的相关系数绝对值最大的变量。如果指定的判据是  $F$  值, 每次将方差分析的  $F$  值最大且大于指定的  $F$  值的变量引入模型。如果指定的判据是大于  $F$  值的概率, 每次将概率最小且小于指定的概率的变量引入模型。

⑤ **Backward**, 向后剔除法。先建立全变量模型。模型中与因变量具有最小偏相关的变量若符合在 **Options** 对话框中所设定的判据被最先从模型中剔除, 然后根据设定的判据, 重复以上步骤, 直到回归方程中不再含有符合剔除判据的自变量为止。

(4) 根据一个设定的变量值, 选择参与回归分析的观测量。将选择变量送入 **Selection Variable** 框中, 单击 **Rule** 按钮, 打开如图 11-4 所示的对话框。

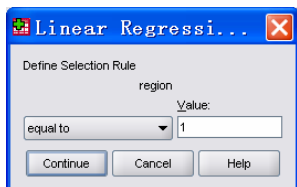


图 11-4 设定运算规则对话框

在下拉列表中选择关系运算法则: **equal to** 等于、**not equal to** 不等于、**less than** 小于、**less than or equal to** 小于等于、**greater than** 大于、**greater than or equal to** 大于等于。然后在 **Value** 框中输入判据, 最后单击 **Continue** 按钮。

(5) 主对话框中, 选择一个变量进入 **Case Label** 框, 其值作为观测量标签。

(6) 选择一个作为权重的变量进入 **WLS Weight** 框中。利用加权最小平方方法给观测量不同的权重值, 它可用来补偿或减少采用不同测量方式所产生的误差。

因变量与自变量, 不能再作为加权变量使用, 加权变量的值如果是零、负数或缺失值, 那么相对应的观测量将被删除。

(7) 单击 **Statistics** 按钮, 打开如图 11-5 的对话框, 选择要输出的统计量。

① **Regression Coefficients** 栏, 有关回归系数的选项。

- **Estimates**, 输出回归系数  $B$ 、 $B$  的标准误、标准化回归系数  $Beta$ 、对回归系数为 0 的假设进行检验的  $T$  值,  $T$  值的双侧检验的显著性概率  $Sig$ 。
- **Confidence intervals**, 输出每一个非标准化回归系数 95% 的置信区间或者一个方差矩阵。
- **Covariance matrix**, 输出非标准化回归系数的协方差矩阵、各变量的相关系数矩阵。

② 与模型拟合及其拟合效果有关的选项。

- **Model fit**, 对拟合过程中引入模型及从模型中剔除的变量, 输出复相关系数及其平方  $R$ 、 $R^2$  及其修正值、估计值的标准误、ANOVA 方差分析表。这是默认选项。
- **R squared chang**, 输出  $R^2_{ch}$ 、 $F_{ch}$ 、 $Sig_{ch}$ 。 $R^2_{ch}$  是当回归方程引入或剔除一个自变量后  $R^2$ 、 $F$  值、 $Sig$  值的变化量。如果较大, 说明进入和从回归方程剔除的可能是一个较好的回归自变量。
- **Descriptives**, 输出合法观测量的数量、变量的平均数、标准差、相关系数矩阵及其单侧检验显著性水平矩阵。

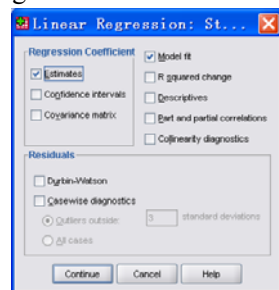


图 11-5 输出统计量对话框

- Part and partial correlations, 输出部分相关系数, 偏相关系数与零阶相关系数。
- Collinearity diagnostics, 输出用来诊断各变量共线性问题的各种统计量和容限值。

③ Residuals 栏, 有关残差分析的选项。

- Durbin-Watson, 输出 Durbin-Watson 统计量以及可能是异常值的观测量诊断表。
- Casewise diagnostics, 输出观测量诊断表。
- Outlines outside standard deviation, 设置异常值的判据, 默认值为 $\geq 3$ 。
- All cases, 输出所有观测量的残差值。

(8) 单击 Plots 按钮, 打开如图 11-6 所示对话框。选择要输出的图形。默认情况下, 不输出图形。

① 在左侧的源变量框中, 任意选择两个变量的组合, 并分别送入 X、Y 轴变量框中。

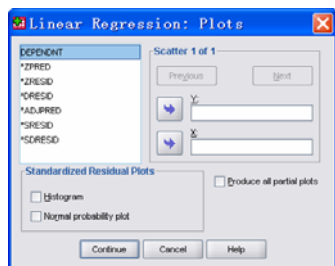


图 11-6 选择图形对话框

可以选择的作图元素有: DEPENDENT 因变量、ZPRED 标准化预测值、ZRESID 标准化残差、DRESID 剔除残差、ADJPRED 修正后预测值、SRESID 学生化残差、SDRESID 学生化剔除残差。

② Standardized Residual Plots 栏, 选择输出标准化残差图。

- Histogram, 输出带有正态曲线的标准化残差的直方图。

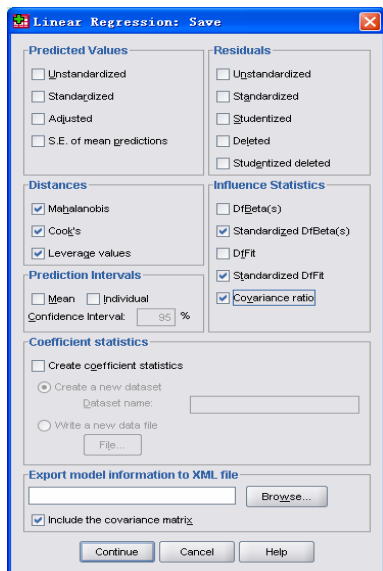


图 11-7 保存变量对话框

• Normal probability plot, 输出 P-P 图, 即残差的正态概率图, 检查残差的正态性。

③ Produce all partial plots, 输出每一个自变量的残差相对于因变量残差的散布图。

(9) 单击 Save 按钮, 打开如图 11-7 所示的 Save 对话框, 指定要保存到数据窗口的新变量。

① Predicted Values 栏, 可选择输出的预测值有: Unstandardized, 非标准化预测值、Standardized, 标准化预测值、Adjusted, 将一个观测值排除在回归方程之外时, 它本身的预测值、S.E. of mean predictions, 预测值的均值标准误。

② Distances 栏, 选择要输出的距离。选择项有: Mahalanobis 距离、Cook 距离、Leverage values 中心点杠杆值。

③ Prediction Intervals 栏, 选择输出预测区间可选项有:

- Mean, 预测区间高低限的平均值。
- Individual, 观测量预测值上、下限的间距。



选择上述两项, 要在 Confidence Intervals 参数框中指定可信区间, 默认 95%, 可输入 0~99.99 之间的值。

④ Residuals 栏, 选择输出的残差有: Unstandardized 非标准化残差、Standardized 标准化残差、Studentized 学生化残差、Deleted 剔除残差、Studentized Deleted 学生化剔除残差。

⑤ Influence Statistics 栏, 输出影响点的统计量

- DfBeta(s), 因排除一个特定的观测值所引起的回归系数的变化值。一般情况下如果此值大于界值  $2/\sqrt{N}$  的绝对值, 则被排除的观测值有可能是影响点。
- Standardized DfBeta(s), 标准化的 DfBeta 值。在数据文件中保存的变量名为 SDBM\_N,  $M=0$  时为常数项,  $M \geq 1$  时为自变量,  $N \geq 1$ , 为  $N$  次运行的模型编号。
- DfFit, 因排除一个特定的观测值所引起的预测值的变化量。
- Standardized DfFit, 标准化的 DfFit 值大于界值  $|2\sqrt{P/N}|$  的观测量可认为是影响点。
- Covariance ratio, 协方差比矩阵。剔除了一个影响点的协方差矩阵与全部观测量的协方差矩阵的比。比值接近 1, 说明观测量对方差矩阵没有显著影响。

⑥ 在 Save to New File 栏, 将回归系数保存到一个指定的文件中。

⑦ 在 Export model information to XML file, 将模型的信息输出到指定的 XML 格式的文件中。单击 Browse 按钮指定保存位置和文件名。

(10) 单击 Option 按钮, 打开如图 11-8 所示的选择项对话框。

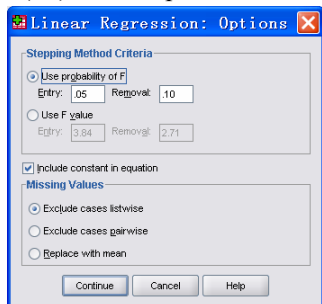


图 11-8 选择对话框

① Stepping Method Criteria 栏, 设置变量引入模型或从模型剔除的判据。

• Use probability of F, 采用 F 检验的概率值作为判据。一个自变量 F 检验 Sig 值  $\leq$  Entry 值时, 该变量被引入回归方程中; 当一个变量 F 检验的 Sig 值  $\geq$  Removal 值时, 该变量从回归方程中剔除。系统默认 Entry=0.05, Removal 为 0.10。可以自定义这两个值。注意必须  $Removal > Entry > 0$ 。增加 Entry 值可使更多变量能够进入方程, 降低 Removal 值可从方程中剔除更多的变量。

- Use F value, 采用 F 值作为判据。系统默认 F 值  $\geq 3.84$  的变量被选入模型中; F 值  $\leq 2.71$  的变量从模型中剔除。可以自定义 Entry、Removal 值。注意, 要  $Entry > Removal > 0$ 。减少 Entry 值可能使更多的变量能够进入方程; 加大 Removal 值可能剔除更多的变量。

② Include constant in equation. 在回归方程中包含常数项, 这是默认选项。

③ Missing Values 栏, 缺失值处理。

- Exclude cases listwise, 将变量表中变量具有缺失值的所有观测量排除在计算之外。
- Exclude cases pariwise, 剔除计算相关系数的一对变量中含有缺失值的观测量。

- Replace with mean, 利用变量的平均数代替缺失值。

### 11.1.6 线性回归分析实例

【例1】使用 data11-13 数据文件, 建立一个以 salbegin 初始工资、prevexp 工作经验、educ 受教育年数为自变量, salary 当前工资为因变量的回归模型。

1. 作数据散点图, 观察因变量与自变量之间关系是否有线性特点。

(1) 按 Graphs→Legacy Dialogs→Scatter/Dot→Simple Scatter 顺序展开作图对话框。

(2) 将变量 salbegin、salary 依次选做 Y 轴变量与 X 轴变量, 单击 OK 按钮。

生成的图形见图 11-9, 其 Y 轴为初始工资, X 轴为当前工资。根据同样操作方法可以作以 salary 为 Y 轴, 分别以其他几个自变量为 X 轴的散点图。

从图中看出初始工资与当前工资存在明显的线性关系, 以初始工作为自变量建立线性回归方程是可能的。对其他可能引入模型的变量, 也应该做出散点图, 有助判断。应当注意, 在最终确定回归方程结果之前还应审查数据中的奇异值、影响点。另外两个变量做对数变换后, 线性关系会更好。读者可以自己作图证明。

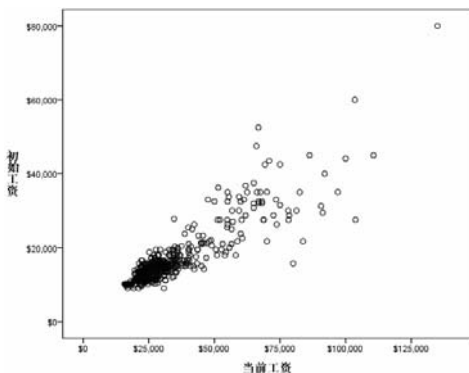


图11-9 初始与当前工资散点图

2. 回归模型的建立

(1) 按 Analyze→Regression→Linear 顺序展开线性模型主对话框。

(2) 在左侧的源变量框中选择变量 salary 作为因变量进入 Dependent 框中。选择变量 salbegin、prevexp、jobtime、educ 作为自变量进入 Independent(s)框中。

(3) 在 Method 框中选择 Stepwise 逐步回归。

(4) 单击 Statistics 按钮, 打开如图 11-5 所示的对话框。选择 Estimates 和 Model fit 输出各种常用统计量; 在 Residuals 栏中选择 Casewise diagnostics 项要求进行奇异值判别; 并在 Outliers outside standard deviation 的参数框中输入 3, 设置观测量标准差大于等于 3 为奇异值; 单击 Continue 按钮返回。

(5) 单击 Save 按钮, 打开如图 11-7 的对话框。选择 Mahalanobis、CooK's、Leverage values、Standardized Dfbeta(s)、Standardized DfFit 和 Covariance ratio 选项, 这些统计量将保存在数据文件中, 用来确定影响点, 单击 Continue 按钮返回。

(6) 为了检测模型的直线性和方差的齐性, 作散点图。单击 Plots 按钮打开 Plots 对话框, 将变量 ZPRED 与 ZRESID 分别选入 X、Y 框中。单击 Continue 按钮返回主对话框。

(7) 单击 OK 按钮, 提交系统执行。

3. 结果输出见表 11-3～表 11-9 及图 11-10～图 11-12。

表 11-3 引入或从模型中剔除的变量

Variables Entered/Removed <sup>a</sup>			
Model	Variables Entered	Variables Removed	Method
1	初始工资		Stepwise (Criteria: Probability of F-to-enter <= .050, Probability of F-to-remove >= .100).
2	工作经验		Stepwise (Criteria: Probability of F-to-enter <= .050, Probability of F-to-remove >= .100).
3	受雇时间		Stepwise (Criteria: Probability of F-to-enter <= .050, Probability of F-to-remove >= .100).
4	受教育时间		Stepwise (Criteria: Probability of F-to-enter <= .050, Probability of F-to-remove >= .100).

a. Dependent Variable: 当前工资

表 11-4 拟合过程小结

Model Summary <sup>a</sup>				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.880 <sup>a</sup>	.775	.774	\$8,115.356
2	.891 <sup>a</sup>	.793	.793	\$7,776.652
3	.897 <sup>a</sup>	.804	.803	\$7,586.187
4	.900 <sup>a</sup>	.810	.809	\$7,465.139

表 11-5 方差分析

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.068E11	1	1.068E11	1622.118	.000 <sup>a</sup>
	Residual	3.109E10	472	6.586E7		
	Total	1.379E11	473			
2	Regression	1.094E11	2	5.472E10	904.752	.000 <sup>a</sup>
	Residual	2.848E10	471	6.048E7		
	Total	1.379E11	473			
3	Regression	1.109E11	3	3.696E10	642.151	.000 <sup>c</sup>
	Residual	2.705E10	470	5.755E7		
	Total	1.379E11	473			
4	Regression	1.118E11	4	2.794E10	501.450	.000 <sup>d</sup>
	Residual	2.614E10	469	5.573E7		
	Total	1.379E11	473			

表11-6 逐步回归过程中不在方程中的变量

Excluded Variables <sup>a</sup>							
Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
					Tolerance	VIF	Minimum Tolerance
1	工作经验 受雇时间	-.137 <sup>a</sup> .102 <sup>a</sup>	-6.558 4.750	.000 .000	-.289 .214	.998 1.000	.998 1.000
2	受教育时间	.172 <sup>a</sup>	6.356	.000	.281	.599	1.669
3	受教育时间	.102 <sup>a</sup>	4.995	.000	.225	1.000	.998
4	受教育时间	.124 <sup>a</sup>	4.363	.000	.197	.520	1.923
5	受教育时间	.113 <sup>a</sup>	4.045	.000	.184	.516	1.937

表 11-3 自左至右各列含义为：拟合步骤编号、每步引入回归方程的自变量、从回归方程中被剔除的自变量、自变量引入或剔除出方程的判据。可以看出，4 个被选择的自变量经过逐步回归过程都进入了回归方程，没有被剔除的变量。

表 11-4 自左至右为：回归方程模型编号、回归方程的复相关系数  $R$ 、 $R^2$ 、修正的  $R^2$ 、估计的标准误。一般随着模型中变量个数的增加， $R^2$  的值也在不断增加，而修正  $R^2$  值与变量的数目无关。本例这个特点不明显。 $R^2$  值的增加这并不意味着模型越好，也未必会减少估计的标准误。修正  $R^2$  值能较确切地反映拟合优度，因此一般从修正  $R^2$  值看拟合优度。除非需要，自变量数量不应太多，多余的自变量会给解释回归方程造成困难。包含多余自变量的模型不但不会改善预测值，反而有可能增加标准误差。由表 11-4 的  $R^2$  以及修正的  $R^2$  值可以看出建立的回归方程比较好。

表 11-5 方差分析表显示回归拟合过程中每一步的方差分析结果。Sig 为  $F$  值大于  $F$  临界值的概率。方差分析结果表明，当回归方程包含不同的自变量时，其显著性概率值均小于 0.001，即拒绝回归系数均为 0 的原假设。因此，最终的回归方程应该包括这 4 个自变量，且方程拟合效果很好。

表11-6显示每一步回归过程中不在方程中的变量信息。

第一步是方程中已经有了一个变量salbegin，外面有3个变量。如果每一个外面的变量单独进入模型，形成两个自变量模型的统计量及检验结果，以及模型中两个自变量之间的共线性诊断。显然，与因变量salary当前工资相关绝对值最高的是工作经验。如果它进入模型，T检验的显著性小于0.001，拒绝回归系数为0的假设。共线性诊断容许度接近1，说明它与第1个进入模型的自变量不具共线性，所以自变量prevexp工作经验第2个进入模型。其他步的分析与此相同。

表 11-7 各步回归过程中的统计量

Coefficients <sup>a</sup>							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1928.206	.888.680	2.170	.031		
	初始工资	1.909	.047	.880	.40276	1.000	1.000
2	(Constant)	3850.718	900.633	4.276	.000		
	初始工资	1.923	.045	.886	.42283	.998	1.002
	工作经验	-22.445	3.422	-.137	-.6558	.998	1.002
3	(Constant)	-10266.629	2959.838	-3.469	.001		
	初始工资	1.927	.044	.888	.43435	.998	1.002
	工作经验	-22.509	3.339	-.138	-.6742	.998	1.002
	受雇时间	173.203	34.677	.102	.4995	1.000	1.000
4	(Constant)	-16149.671	3255.470	-4.961	.000		
	初始工资	1.768	.059	.815	.30111	.551	1.814
	工作经验	-17.303	3.528	-.106	-.4904	.865	1.156
	受雇时间	161.486	34.246	.095	.4715	.992	1.008
	受教育时间	669.914	165.596	.113	.4045	.516	1.937

a. Dependent Variable: 当前工资

注意：表中给出的是中文变量标签。以下模型中的变量均为变量名。

表 11-7 每一步回归过程的统计量及检验结果。

$B$  为非标准化回归系数，也称偏回归系数，它是在控制了其他变量之后得到的。只有当所有的自变量单位统一时，它们才有可比性。由方差分析得出回归方程有统计意义，而回归方程中的每一个偏回归系数不一定都有显著性，但至少要有一个是显著的。

$Beta$  为标准化的回归系数，它具有可比性。是所有的变量按统一方法标准化后拟合的回归方程中各标准化变量的系数。

$t$  为偏回归系数为 0（和常数项为 0）的假设检验的  $t$  值，具有较好预测效果的变量的  $t$  值应大于 2 或者小于 -2。Sig 为假设检验的显著性概率。4 步回归的各变量和常数项的检验的  $p$  值均小于 0.05。

共线性统计量给出了容许度值和方差膨胀因子。

以第 2 步为例说明这些统计量：回归方程中包含常数项 (Constant) 和自变量  $salbegin$ 、 $prevexp$ ；因变量为  $salary$ 。共线性诊断的指标：容忍度 (Tolerance) 分别为 0.998、0.998，接近 1，方差膨胀因子 VIF 值都不大，可以认为两个自变量之间不存在共线性问题。

模型 2： $salary = 3850.7 + 1.92salbegin - 22.4prevexp$ 。方程常数项和两个自变量  $T$  检验的显著水平值均小于 0.001，拒绝常数项和回归系数为 0 的假设。方程成立。

模型 5： $salary = -16149.7 + 1.77salbegin - 17.30Prevexp + 161.48jobtime + 669.9educ$ 。是最后的回归模型。每个自变量的显著水平值都小于 0.001。各个自变量的容许度值分别为 0.551、0.865、0.992、0.516 没有出现特别小的数值，相应的 VIF 值分别为 1.814、1.156、1.008、1.937，没有很大的数值出现，说明方程中各自变量之间没有出现共线性问题。

表 11-8 为异常值诊断表。自左至右：奇异值观测编号、标准化残差、因变量当前工资的值、

表 11-8 当前工资变量的异常值表

Casewise Diagnostics <sup>a</sup>				
Case Num <sup>***</sup>	Std. Residual	Current Salary	Predicted Value	Residual
18	6.173	\$103,750	\$57,671.26	\$46,078.744
103	3.348	\$97,000	\$72,009.89	\$24,990.108
106	3.781	\$91,250	\$63,026.82	\$28,223.179
160	-3.194	\$66,000	\$89,843.83	\$-23,843.827
205	-3.965	\$66,750	\$96,350.44	\$-29,600.439
218	6.108	\$80,000	\$34,405.27	\$45,594.728
274	5.113	\$83,750	\$45,581.96	\$38,168.038
449	3.590	\$70,000	\$43,200.04	\$26,799.959
454	3.831	\$90,625	\$62,027.14	\$28,597.858

a. Dependent Variable: Current Salary

当前工资的预测值、残差。表中给出了被怀疑为异常值的观测量的编号，这些观测量之所以被怀疑为异常值，是因为它们的标准化残差绝对值都大于设置值 3。

表 11-9、图 11-10 和图 11-11 配合可查找影响点。表 11-9 中的 Mahal. Distance、Cook's Distance、Centered Leverage Value 统计量值，都可以帮助判断是否含有影响点。以 Mahal. Distance 为例，其值范围越大，越可能含有影响点。

表 11-9 残差统计量

Residuals Statistics <sup>a</sup>					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	\$13,354.82	\$150,076.77	\$34,419.57	\$15,372.742	474
Std. Predicted Value	-1.370	7.524	.000	1.000	474
Standard Error of Predicted Value	391.071	3191.216	721.093	260.806	474
Adjusted Predicted Value	\$13,290.94	\$153,447.97	\$34,425.45	\$15,451.094	474
Residual	\$-29,600.439	\$46,078.746	\$-1.194E-12	\$7,433.507	474
Std. Residual	-3.965	6.173	.000	.996	474
Stud. Residual	-4.089	6.209	.000	1.004	474
Deleted Residual	\$-31,485.213	\$46,621.117	\$-5.882	\$7,553.608	474
Stud. Deleted Residual	-4.160	6.474	.002	1.016	474
Mahal. Distance	.300	85.439	3.992	5.306	474
Cook's Distance	.000	.223	.003	.016	474
Centered Leverage Value	.001	.181	.008	.011	474

a. Dependent Variable: Current Salary

	MAH_1	COO_1	COV_1	SDF_1
28	3.05908	.00005	1.01914	-.01635
29	85.43873	.22320	1.17231	-1.0609
30	3.02689	.00019	1.01821	-.03072
31	2.63989	.00070	1.01367	.05909
32	15.38601	.05906	.95831	.54764
33	2.84042	.00156	1.00871	.08818
34	11.30935	.01472	1.00761	.27178

图11-10 判定影响点的各种常用统计量

	SDB0_1	SDB1_1	SDB2_1	SDB3_1	SDB4_1
29	.14326	-1.0054	.11259	-.22739	.48914
30	.02116	.01126	.00131	-.02152	-.01166
31	-.0250	.00466	-.0098	.04701	-.01787
32	-.2404	.41245	-.0205	.21563	-.07845
33	-.0631	-.02912	.00103	.06343	.03278
34	-.1561	.15770	.05034	.11722	.01547

图11-11 标准化回归系数变化量

在图 11-10 中根据前述的判据，可以大致断定 29、32 号观测量为影响点。但是，进一步观察后发现 34 号观测量也存在是影响点的可能性，为判别某一观测量是否为影响点，可以比较此观测量在与不在回归方程中时，标准化回归系数的差异。如图 11-11 所示。

图 11-11 是数据文件中的新变量。SDB0\_1~SDB4\_1 分别对应常数项和第 1~第 4 个自变量的标准化回归系数的变化量。第 34 号观测量的 SDB0\_1、SDB1\_1、SDB3\_1 的值大于  $2/\sqrt{N}$ （本例  $N=474$ ，值为 0.09186），即常数项、第 1、3 个自变量 jobtime、educ 标准化回归系数变异较大。以此初步认定 34 号观测量也为影响点。当然对影响点的判断仅仅凭借一个指标往往是不充分的，还需要用其他指标进行比较判断。

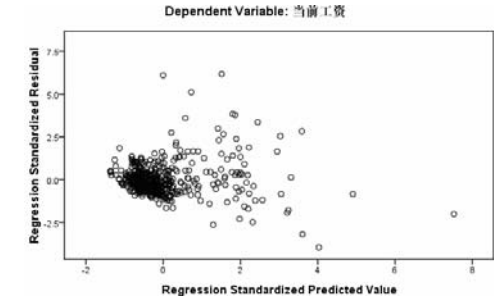


图 11-12 当前工资的预测值与学生化残差散点图

图11-12为当前工资的标准化预测值与其学生化残差散点图中可以看到绝大部分观测

量随机地落在垂直围绕 $\pm 2$ 的范围内, 预测值与学生化残差值之间没有明显的关系, 所以回归方程应该满足线性与方差齐性的假设且拟合效果较好。

【例 2】为说明共线性的诊断的方法, 特举一全模型的例题。

打开线性模型主对话框, 选择变量 salary 作为因变量, 选择 salbegin、prevexp、jobtime、educ、age 作为自变量。在 Method 框中选择 Enter 方式。在 Statistics 对话框中选择 Collinearity diagnostics, 要求进行共线性诊断, 单击 OK 进行分析。结果如表 11-10 所示。

表 11-10 共线性诊断指标

Collinearity Diagnostics <sup>a</sup>									
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	受教育时间	初始工资	受雇时间	工作经验	年龄
1	1	5.313	1.000	.00	.00	.00	.00	.00	.00
	2	.509	3.232	.00	.00	.01	.00	.30	.00
	3	.136	6.243	.01	.00	.49	.01	.03	.01
	4	.022	15.706	.00	.51	.34	.00	.46	.36
	5	.014	19.226	.00	.32	.13	.49	.14	.40
	6	.006	30.287	.99	.17	.02	.49	.07	.23

a. Dependent Variable: 当前工资

Coefficients <sup>a</sup>			
Model		Collinearity Statistics	
		Tolerance	VIF
1	受教育时间	.508	1.967
	初始工资	.551	1.815
	受雇时间	.983	1.018
	工作经验	.347	2.882
	年龄	.347	2.880

a. Dependent Variable: 当前工资

(a)

(b)

表 11-10(a) Eigenvalue 特征值一列中有 3 个特征值分别为 0.022、0.014、0.006, 都非常接近 0, 对应的 3 个条件指数分别为 15.706、19.226、30.287 都大于 15, 两个指标都说明在这三个变量间可能存在共线性。条件指数大于 30 一定存在严重的共线性。

表 11-10(b) 模型中的变量、容许度和方差膨胀因子。该表显示变量工作经验与年龄的方差膨胀因子值分别为 2.882、2.880 全部大于 2, 也说明存在共线性问题。

建议对可能存在共线性的变量做偏相关, 进一步分析在哪两个变量之间存在相关关系, 选择代表性变量参与回归。

## 11.2 曲线估计

### 11.2.1 曲线回归概述

#### 1. 一般概念

线性回归不能解决所有的问题。尽管有可能通过一些函数的转换, 在一定范围内将因、自变量之间的关系转换为线性关系, 但这种转换有可能导致更为复杂的计算或失真。

SPSS 提供了 11 种不同的曲线回归模型中。如果线性模型不能确定哪一种为最佳模型, 可以试试选择曲线拟合的方法建立一个简单而又比较合适的模型。

#### 2. 数据要求

(1) 自变量与因变量应该是数值型变量。如果自变量以时间间隔测度, 要求因变量也是以时间间隔测度的变量, 而且因、自变量使用的时间间隔和单位应是完全相同的。

(2) 模型的残差应该呈正态分布。如果选择了线性模型, 因变量必须是正态分布的,

且所有的观测量应该是独立的。

11.2.2 曲线回归过程

(1) 按 Analyze→Regression→Curve Estimation 顺序打开如图 11-13 所示的对话框。

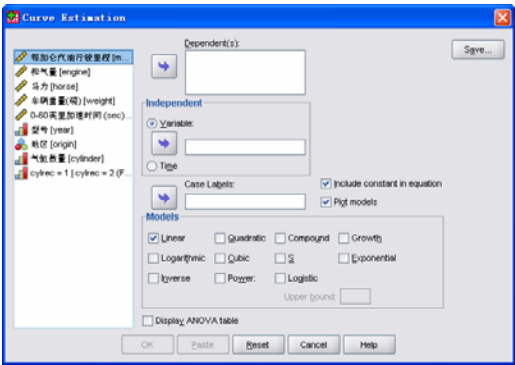


图 11-13 曲线估计对话框

(2) 在源变量框中选择一个或多个因变量，送入 Dependent(s)框中。

(3) 在左侧的源变量框中选择自变量，送入 Independent 的 Variable 框中。如果因变量是时间间隔测度的，直接指定时间选项 Time。

(4) 在 Models 栏中选择一个或多个拟合模型。各模型解释如表 11-11 所示。

在各公式中， $t$  表示时间或所指定的自变量， $b_0$  为常数项。 $b_n$  为自变量第  $n$  次项的回归系数。 $\ln$  是以  $e$  为底的自然对数。

如果选择了 Logistic 模型，模型中的  $u$  值必须是大于因变量中的最大值的正数。在 Upper bound 框中输入这一上限值。

(5) 根据需要进行以下选项：

- ① Include constant in equation，包括常数项。
- ② Display ANOVA table，进行方差分析并输出方差分析表。
- ③ Plot models，产生模型图。
- ④ 在源变量框中选择作为标识观测量的变量，送入 Case Labels 框中。

(6) 单击 Save 按钮，打开对话框，如图 11-14 所示。选择要保存在数据文件中的新变量。其中包括预测值、残差、预测区间、显著性水平等变量。系统默认的新变量名与说明显示在输出窗口中。

表 11-11 不同模型的表示

模型名称	回归方程	相应的线性回归方程
Linear	$y=b_0+b_1t$	
Quadratic	$y=b_0+b_1t+b_2t^2$	
Compound	$y=b_0(b_1^t)$	$\ln(y)=\ln(b_0)+[\ln(b_1)]t$
Growth	$y=e^{(b_0+b_1t)}$	$\ln(y)=b_0+b_1t$
Logarithmic	$y=b_0+b_1\ln(t)$	
Cubic	$y=b_0+b_1t+b_2t^2+b_3t^3$	
S	$y=e^{b_0+b_1/t}$	$\ln(y)=b_0+b_1/t$
Exponential	$y=b_0e^{bt}$	$\ln(y)=\ln(b_0)+b_1t$
Inverse	$y=b_0+(b_1/t)$	
Power	$y=b_0(t^{-b_1})$	$\ln(y)=\ln(b_0)+b_1\ln(t)$
Logistic	$y=1/(1/u+b_0(b_1^t))$	$\ln(1/y - 1/u)=\ln(b_0+[\ln(b_1)t])$

① Save Variables 栏，保存变量。选项有：Predicted values 因变量的预测值；Residuals 残差值；Prediction intervals 预测区间；在 Confidence interval 框中设置预测值的可信区间。

② Predicted cases 栏，预测观测量。如果自变量为时间变量，可以在该栏中指定一

种超出当前数据时间序列范围的预测周期。

- Predict from estimation period through last case, 使用预先设定好的, 求出估计周期到最后应该观测量的的预测值。估计周期和预测范围可以通过 Data 菜单中的 Select case 命令设置。如果没有预先设置估计周期, 计算时使用所有的观测值。
- Predict though, 根据预先设定的周期, 对特定的数据、在指定的时间内进行预测。如果预测值的范围超出了时间序列的范围, 应该选择该项, 并在随后的 Observation 框中输入预测周期的末端值。

(7) 单击 OK 按钮提交运行,

在大多数情况下, 对变量之间关系的认识往往模糊不清。需要先绘制散点图, 根据数据分布特点, 确定应采用的模型。可以多指定几个模型进行拟合, 根据输出的统计量, 例如  $R^2$  值, 结合图形综合考虑, 确定最佳模型。

### 11.2.3 曲线回归分析实例

【例 3】用 data11-01 中的数据研究车重 weightt 与每加仑公里数 mpg 之间的关系。

1. 制作观测量数据的散点图并初步选择模型。

打开数据文件, 按 Graphs→Legacy Dialogs→Scatter/Dot 顺序打开做散点图的对话框。

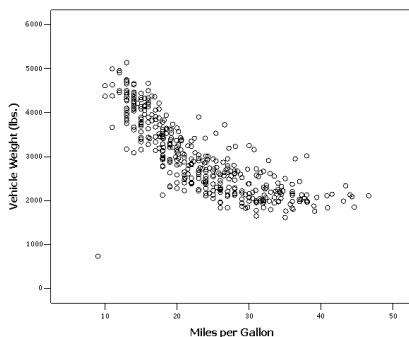


图 11-15 每加仑里程与车重散点图

以变量 mpg 做 X 轴, 变量 weightt 做 Y 轴, 得到散点图 11-15。可以看到, 两个变量间呈现明显的曲线关系。

2. 建立若干个曲线模型进行比较

(1) 按 Analyze → regression → Curve Estimation 顺序打开主对话框。选择变量 mpg 作为因变量, weight 作为自变量。

(2) 在 Model 框中选择 Quadratic 二次、Cubic 三次与 Compound 指数模型。

(3) 选择 Display ANOVA table、Plot model 及 Including constant in equation 选项, 要求输出

方差分析的结果和模型图形, 方程包括常数项。单击 OK 按钮, 提交系统执行。

3. 输出结果见表 11-12~表 11-14, 以及图 11-16。每组表对应 1 个模型的输出, 每组表格含 3 个子表: model summary 模型摘要、ANOVA 方差分析、Coefficients 系数表。

4. 结果分析

Model Summary 表列出复相关系数  $R$ 、判定系数  $R^2$ 、 $R^2$  的修正值、标准误。

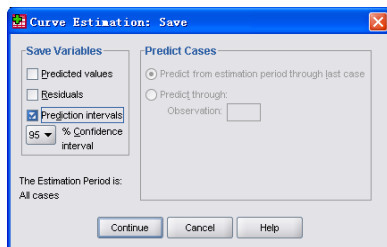


图 11-14 保存对话框



ANOVA 为方差分析结果：二次模型的  $F$  值为 377.209，三次模型的  $F$  值为 286.476；指数模型的  $F$  值为 957.936， $p$  值（表中 Sig）小于 0.0001。三个回归方程均有统计意义。

表 11-12 Compound 结果

Model Summary			
R	R Square	Adjusted R Square	Std. Error of the Estimate
.841	.708	.707	.184

The independent variable is 车辆重量(磅).

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	32.405	1	32.405	957.936	.000
Residual	13.396	396	.034		
Total	45.800	397			

The independent variable is 车辆重量(磅).

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
车辆重量(磅)	.099966	.000	.431	9.204E4	.000
(Constant)	60.152	2.013		29.887	.000

The dependent variable is ln(每加仑汽油行驶里程).

表 11-13 Quadratic 结果

Model Summary			
R	R Square	Adjusted R Square	Std. Error of the Estimate
.810	.656	.655	4.593

The independent variable is 车辆重量(磅).

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	15918.130	2	7959.065	377.209	.000
Residual	8334.445	395	21.100		
Total	24252.575	397			

The independent variable is 车辆重量(磅).

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
车辆重量(磅)	-.012	.002	-.130	-6.094	.000
车辆重量(磅)**2	7.597E-7	.000	.528	2.419	.016
(Constant)	52.540	3.030		17.337	.000

表 11-14 Cubic 结果

Model Summary			
R	R Square	Adjusted R Square	Std. Error of the Estimate
.828	.686	.683	4.399

The independent variable is 车辆重量(磅).

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	16629.063	3	5543.021	286.476	.000
Residual	7623.513	394	19.349		
Total	24252.575	397			

The independent variable is 车辆重量(磅).

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
车辆重量(磅)	.033	.008	3.598	4.286	.000
车辆重量(磅)**2	-1.434E-5	.000	-9.968	-5.715	.000
车辆重量(磅)**3	1.591E-9	.000	5.655		
(Constant)	9.555	7.662		1.247	.213

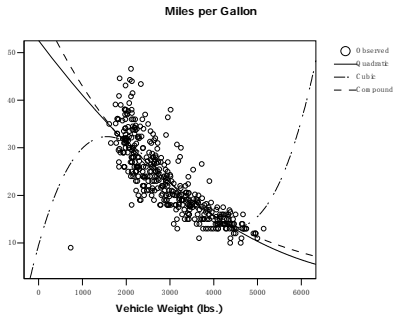


图 11-16 三种模型的图形

Coefficients 中显示回归系数 B、标准化回归系数 Beta 及其检验结果，由此得出各种模型的回归模型如下：

二次： $\text{mpg}=52.540-0.012\times\text{weight}+7.597\times10^{-7}\times\text{weight}^2$

三次： $\text{mpg}=9.555+0.033\times\text{weight}-1.434\times10^{-5}\times\text{weight}^2+1.591\times10^{-9}\times\text{weight}^3$

指数模型： $\text{mpg}=60.15\times0.9996^{\text{weight}}$

图 11-16 是三种模型的图形，似乎虚线即指数曲线对观测量的拟合稍好一些。图形只是对模型的取舍起辅助作用，最终的模型判定还是要通过对统计量的分析与研究进行。

① 比较三个模型的修正  $R^2$  值。指数模型的  $\text{Adjusted } R^2=0.708$  最大，三次模型次之， $R^2=0.686$ ，二次模型的  $R^2=0.656$  最小。由此可以判断，拟合最好的是指数模型。

② 方差分析的  $F$  值概率均小于 0.001，因此比较  $F$  值。指数模型的  $F=957.936$  最大，二次模型次之， $F=377.209$ ，三次模型的  $F=286.476$  最小。

通过以上判断得出最佳模型为： $\text{mpg}=60.15\times0.9996^{\text{weight}}$ 。

注意：输出窗中表格中的小数显示位数设置为 5。

## 11.3 二项逻辑斯谛回归

在现实世界中,经常需要判断一些事情是否将要发生,候选人是否会当选等。这类问题的特点是因变量只有两个值,发生(是)或者不发生(否)。这就要求建立的模型必须保证因变量的取值是0、1。可是,大多数模型的因变量值常常处于一个实数集中,与因变量只有两个值的条件相悖。

本节介绍一种对因变量数据假设要求不高,并且可以用来预测具有两分特点的因变量概率的统计方法:二项逻辑斯谛(Binary Logistic)回归模型。

当因变量具有两类以上的分类时,可以参考9.4节的多分变量的Logistic回归。

### 11.3.1 Logistic回归模型

#### 1. Logistic 模型

在逻辑斯谛回归中可以直接预测观测量相对于某一事件的发生概率,如果只有一个自变量,回归模型可以写做

$$Prob(event) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$

其中  $b_1$  和  $b_0$  为自变量  $x$  的系数和常数,  $e$  为自然数。其曲线如图 11-17 所示。包含一个以上自变量的模型为

$$Prob(event) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

其中:  $z = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$  ( $p$  为自变量的数量) 某一件事情不发生的概率为

$$Prob(\text{no event}) = 1 - Prob(event)$$

逻辑斯谛模型的建立使用最大似然比法和迭代方法。

#### 2. 数据要求

(1) 因变量应具二分特点, 自变量可以是分类变量或等间隔测度的变量。如果自变量是分类变量, 应为二分变量或被重新编码为指示变量。指示变量有两种编码方式。

① 指示变量编码方案。例如, 当分类变量有三个水平(高、中、低), 就要创建两个新的指示变量。第一个变量: 1 为低水平, 0 为其他水平; 第二个变量: 1 为中间水平值, 0 为其他水平值; 高水平观测量的两个变量值同时为 0。哪种水平为 0 值可任意决定。表 11-17 中参考类别的系数为 0。使用指示变量编码方法, 只能比较每一类与参考类之间的效应差异。如果要比较每一类与整体的综合效果, 应该选择如表 11-15 所示的编码方式。

② 背离编码方案。与编码方法一的区别仅仅在于新变量中最后一类被赋予-1 的编

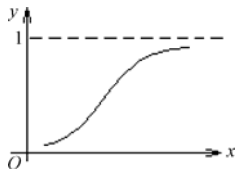


图 11-17 Logistic 回归曲线

码值。利用这种编码方法，逻辑斯谛回归系数展示每一类与各类综合效果的差异。参见表 11-16，对于每一个 SPSS 创建的新变量，其系数代表了与综合效果之间的差异。注意，最后一类的值应该是前两种系数之和并取负值。

表 11-15 指示变量编码方法

表 11-16 背离编码方法

Varvale (变量值)		Frequency (频数)	Parameter coding (指示变量的编码设置)	
			(1)	(2)
Catacid (变量名称)	1.00	15	1	0
	2.00	20	0	1
	3.00	18	0	0

Varvale (变量值)		Frequency (频数)	Parameter coding (指示变量的编码设置)	
			(1)	(2)
Catacid (变量名称)	1.00	15	1	0
	2.00	20	0	1
	3.00	18	-1	-1

(2) 自变量数据最好为多元正态分布，自变量间的共线性会导致估计偏差。当观测量分组完全依据分组变量时，这个方法十分有效；当观测量分组依据某连续型数值时（如根据智商得分可分高智商、低智商），此方法会丢失连续型数据的信息，应考虑线性模型。

3. Logistic 回归系数

为了理解 Logistic 回归系数的含义，可以将回归方程改写为某一事件发生的比率，一个事件的比率被定义为它发生的可能性与不发生的可能性之比。例如，抛一枚硬币后，其正面向上的比率为  $0.5/0.5=1$ ，从 52 张牌中抽出一张 A 的几率为  $(4/52)/(48/52)=1/12$ ，这里不要将几率的含义与“概率”混淆，其概率值为  $4/52=1/13$ 。

首先把 Logistic 方程写作几率的对数，命名为 Logit。

$$\log\left(\frac{Prob(event)}{Prob(no\ event)}\right) = b_0 + b_1x_1 + \Lambda + b_px_p$$

可以看出，逻辑斯谛方程的回归系数可以解释为一个单位的自变量的变化所引起的比率的对数的改变值。由于理解几率要比理解几率的对数容易一些，所以将逻辑斯谛方程式写为

$$\frac{Prob(event)}{Prob(no\ event)} = e^{b_0 + b_1x_1 + \Lambda + b_px_p}$$

当第  $i$  个自变量发生一个单位的变化时，比率的变化值为  $\text{Exp}(b_i)$ 。自变量的系数为正值，意味着事件发生的比率会增加， $\text{Exp}(b_i)$  的值大于 1；如果自变量的系数为负值，意味着事件发生的比率将会减少，此值小于 1；当  $b_i$  为 0 时，此值等于 1。

4. 评价模型

建立模型后，需要判断拟合的优劣。对大样本量的数据，最好将数据分成两部分，用一部分数据建立回归方程，再将另一部分数据带入方程，评定模型对数据的拟合情况。

(1) 系数检验

对于较大样本的系数检验，使用基于卡方分布的 Wald 统计量。当自由度为 1 时，Wald 值为变量系数与其标准误比值的平方。对于两类以上的分类变量来说，Wald 统计量为  $W=B'V^{-1}B$ ，此处  $B$  为分类变量系数的极大似然估计向量， $V^{-1}$  为变量系数渐近方

差-协方差矩阵的逆矩阵。

Wald 统计量的弱点是当回归系数的绝对值变大时, 其标准误将发生更大的改变, Wald 值就会变得很小, 导致拒绝回归系数为 0 的零假设失败, 即认为变量的回归系数为 0。因此当变量的系数很大时, 不应该依据 Wald 进行检验, 应该建立包含与不包含要检验的变量的两个模型, 利用对数似然比的变化值进行检验。可以选择 Backward LR 方式作为变量的选择方法。

## (2) 模型判别和模型校验

① 模型判别, 依据对事件发生的可能性的估计, 评估模型区分两组数据的能力。好的模型会将高概率的数值赋值给经常发生事件的观测量, 不大可能发生的事件观测量得到较小的概率值, 两种数据的概率不会发生重叠。

经常用来检查模型“判别”能力的指标为  $C$  统计量, 其值的范围为从 0.5~1。0.5 表示模型对观测量的类别“判别”作用非常弱, 1 表示强判别力。

SPSS 的逻辑斯蒂回归过程, 先计算预测概率, 再利用 ROC 功能计算  $C$  统计量。

② 模型校验, 评估观测概率、预测概率与整个概率之间的关系, 它对观测量概率与预测概率之间的差异进行解释。当协变量配对的数量巨大, 且不能使用标准拟合度卡方检验时, 常用的检测方法 Hosmer 和 Lemeshow 卡方统计量非常有效。

计算 Hosmer-Lemeshow 卡方统计量, 先计算每一组中事件发生的实际观测数量与预测数量之间的差异, 然后按  $(\text{观测数量} - \text{预测数量})^2 / \text{预测数量}$  计算, 卡方值为各分组中此值的和。

实际操作方法是根据估计观测数量的预测概率将观测量分成数量大致相同的 10 个组, 观察观测到的数量与预测发生事件的数量以及预测不发生事件的数量之间的比较结果。卡方检测用来评价实际事件发生与预测事件发生之间的数量差别。使用这种鉴别方法时数据要相当大, 以确保在大多数组别中至少有 5 个以上的观测量, 同时所有的组别的预测值大于 1。

Hosmer 和 Lemeshow 卡方统计量的结果很大程度上与观测量的分组情况有关。如果分组数很小, 得出的结果很可能与实际情况不符。但如果观测量数量很多, Hosmer 和 Lemeshow 卡方统计量的结果也会变大。因此虽然 Hosmer 和 Lemeshow 卡方统计量在进行“模型校对”检测时是一种非常有效的方法, 但必须结合观测量进行解释。

(3) 模型的拟合度是判别模型与样本的拟合优劣。利用已有的参数, 得出的观测结果的可能性称为“似然比”。似然比的值小于 1, 习惯上用对数似然比值乘以 -2 来度量模型对数据的拟合度, 记做  $-2\ln$ 。好的模型的似然比值较高, 其  $-2\ln$  值相对较小 (如果模型 100% 的完美, 似然比值等于 1,  $-2\ln$  值为 0)。似然比值的变化说明当变量进入与被剔除出模型时模型对数据拟合度方面的变化。

常用的 3 种卡方统计量分别为 Model、Block 和 Setp。

① Model 统计量检验除常数项以外, 模型中所有变量系数为零的假设。卡方值为当

前模型的与模型中只包含常数项的-2ll-likelihood 之差。

② Block 卡方值为当前模型与后一组变量进入模型后的-2ll-likelihood 值之差。如果选择了多组变量,那么 Block 卡方值用来对最后一组变量系数为 0 的零假设进行检验。

③ Step 卡方值是在建立模型的过程中,当前与下一步-2ll 之间的差值。它用来对最后一个加入模型的变量系数为 0 的零假设进行检验。

(4) 评价包含所有变量模型的拟合效果。

① Cox & Snell  $R^2$  和 Nagelkerke  $R^2$  统计量,与线性模型中的  $R^2$  相似。是对逻辑斯谛模型变异中可解释部分的量化。

$$\text{Cox \& Snell } R^2 = 1 - \left( \frac{L(0)}{L(B)} \right)^{\frac{2}{N}},$$

式中  $L(0)$  为方程中只包含常数项时的似然比值,  $L(B)$  为方程包含设定变量时的似然比值,  $N$  为样本量。Cox & Snell  $\tilde{R}^2$  统计量最大值不可能为 1, 1991 年 Nagelkerke 修改了 Cox & Snell  $R^2$  统计量,使得 Nagelkerke  $R^2$  的最大值可以为 1。

$$\text{Nagelkerke } \tilde{R}^2 = \frac{R^2}{R_{\max}^2}$$

式中  $R_{\max}^2 = 1 - (L(0))^{\frac{2}{N}}$ 。反映由回归方程解释的变异百分比。

② 偏差,对于每一个观测量,其偏差值为  $(-2\log(\text{预测概率}))^{0.5}$ 。例如,某男性患者预测其没有患恶性淋巴结的概率为 0.80 时,其偏差为  $-\sqrt{-2\log(0.8)} = -0.668$ 。大样本数据的偏差往往近似正态分布。偏差较大暗示模型拟合数据欠佳。学生化残差与偏差之差可以用来检测非常态数据。以  $P_i$  为第  $i$  个观测量的预测概率,残差的 Logit 值计算公式为

$$\frac{\text{residual}_i}{P_i(1-P_i)}$$

(5) 影响点的查找

① 杠杆值 (Leverage) 检测哪些观测量对预测值产生影响较大。与线性回归不同,在逻辑斯谛回归中杠杆值依据因变量得分和设计矩阵。其值在 0~1 之间,它们的均值为  $P/N$ ,其中  $P$  为模型中估计参数的个数 (包括常数项),  $N$  为观测量的个数。对于那些预测概率值大于 0.9 或小于 0.1 观测量来说,虽然观测量具有影响力,但其杠杆值较小。

② Cook 距离用来检测观测量的影响力。说明如果删除了一个观测量后对这个观测量残差的影响和对其他观测量残差的影响。

Cook 距离为  $D_i = \frac{Z_i^2 \times h_i}{(1-h_i)}$ , 式中  $Z_i$  为标准化残差,  $h_i$  为杠杆值。

③ DfBeta 统计量,即当删除一个观测量后逻辑斯谛系数的变化值。

$\text{DfBeta}(b_1^{(i)}) = b_1 - b_1^{(i)}$ , 其中  $b_1$  为当所有观测量包括在模型中时的系数值,  $b_1^{(i)}$  为

排除第  $i$  个观测量后的系数值。较大的变化值暗示应对此观测量给予重新检查。

(6) 与线性回归相同, 交互项可以作为新变量参与回归分析并包含在回归方程中。

### 11.3.2 二项逻辑斯谛回归过程

1. 按 Analyze→Regression→Binary Logistic 顺序打开如图 11-18 所示的对话框。

2. 选择一个具有两分属性的变量作为因变量送入 Dependent 框。

3. 选择一个或多个变量为协变量送入 Covariates 框。也可以同时选择两个和多个变量作为交互项, 单击 “>a\*b>” 按钮, 将它们送入 Covariates 框。

4. 在 Method 选项框中确定一种自变量进入模型的方式。

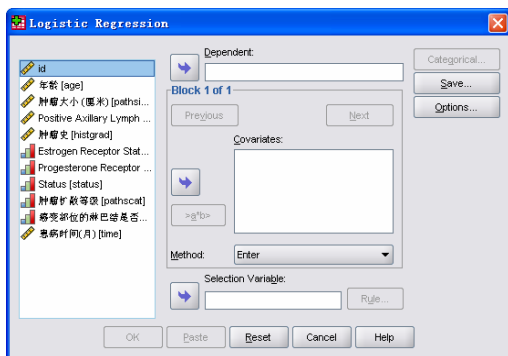


图 11-18 二项逻辑斯谛过程对话框

(1) Enter, 自变量全部进入模型。

(2) Forward:conditional, 向前逐步选择法。将变量剔除出模型的依据是, 条件参数估计的似然比统计量的概率值。

(3) Forward:LR, 依据最大偏似然估计所得的似然比统计量的概率值, 向前逐步选择。

(4) Forward:Wald, 依据 Wald 统计量的概率值向前逐步选择变量。

(5) Backward:conditional, 根据条件参数估计似然比统计量的概率值, 向后逐步剔除。

(6) Backward:LR, 依据最大偏似然估计值统计量的概率值向后逐步剔除。

(7) Backward:Wald, 根据 Wald 统计量的概率向后逐步剔除。

5. Selection Variable 框, 根据指定变量的取值范围, 确定参与分析的观测量。在源变量框中选择一个变量, 送入 Selection Variable 框中。

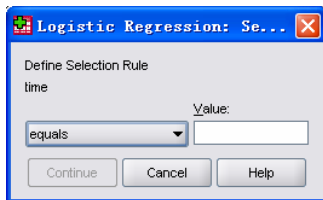


图 11-19 设定选择规则对话框

单击 Rule 按钮, 打开 Setrule 对话框, 如图 11-19 所示, 设置选择观测量的标准。例如, 要选择 time 等于 100 秒的变量, 那么选择 time 变量后在算数操作符框中选择 equals, 然后在 Value 框中输入 100。

SPSS 会将选择的观测量与非选择的观测量的计算结果全部显示出来。

6. 单击 Categorical 按钮, 展开如图 11-20 对话框。设置处理分类变量的方式。

(1) 在 Covariates 框中包含了在主对话框中已经选择好的全部协变量及交互项。

(2) Categorical Covariates 框中列出了所选择的分类变量, 其后面的括号中显示的是各组间的对比方案, 字符串变量将自动进入 Categorical Covariates 框。

(3) Change Covariates 栏, 设置分类协变量中各类水平的对比方式。

① Indicator, 指示出是否同属于参考分类, 参考分类在对比矩阵中以一横排 0 表示。

② Simple, 每一种分类的预测变量 (参考类别外) 效应都与参考类别效应进行比较。

③ Difference 选项, 除第一类外, 每类的预测变量效应都与其前所有各分类的平均效应进行比较, 也称作逆 Helmert 对比。

④ Helmert, 除最后一类外, 每类的预测变量效应都与其后所有各类的平均效应进行比较。

⑤ Repeated, 除第一类外, 每类的预测变量效应都与其前一种分类的效应进行比较。

⑥ Polynomial, 对角多项式对比, 要求每类水平相同, 仅适用于数字型变量。

⑦ Deviation, 每类的预测变量 (参考分类除外) 效应与总体效应进行比较。

⑧ Reference Category, 如果你选择了 Deviation、Simple、Indicator 对比方式, 可选择 First 或 Last, 指定分类变量的第一类或最后一类作为参考类。

如果改变了 Change Covariates 的设置, 单击 Change 按钮以示对选项的确定。

7. 单击 Save 按钮, 展开如图 11-21 所示的对话框, 选择在数据窗中保存的新变量。

(1) Predicted Values 预测值栏, Probabilities, 新变量是每个观测量发生特定事件的预测概率; Group Membership, 依据预测概率得到的每个观测量的预测分组。

(2) Influence 栏, 每一个观测量的影响力指标。包括 Cook 距离、杠杆值 Leverage values 和 DfBeta 统计量。

(3) Residuals 栏, 残差: Unstandardized 非标准化残差、Logit 残差、Studentized 学生化残差、Standardized 标准化残差和 Deviance 偏差。

(4) Export model information to XML file 栏指定输出模型信息到 XML 格式的文件, 单击 Brows 按钮, 确定保存位置和文件名。选择 Include covariance matrix 输出还包括协方差矩阵。

8. 单击 Options 按钮, 展开如图 11-22 所示的 Options 对话框, 设置各种检测参数。

(1) Statistics and Plots 栏, 选择要求输出的统计量与图表。

① Classification Plots, 因变量的预测值与观测值的分类直方图。

② Hosmer-lemeshow goodness-of-fit, 拟合良好度统计量。

③ Casewise listing of residuals, 对每个观测量输出非标准化残差、预测概率、观测量的实际与预测分组水平。

- Outliers outside \_\_ std. Dev, 在空格处输入一个正数, 表示要求只输出那些标准化残差值大于输入值的观测量的各种统计量。

- All cases, 输出所有观测量的各种统计量。

- Correlations of estimates, 输出方程中各变量估计参数的相关系数矩阵。

- Iteration history, 进行参数估计时, 每一步迭代输出的相关系数和对数似然比值。

- CI for exp(B), 在此处输入 1~99 的数值。

(2) Display 栏设置输出范围: At each step, 对每步计算过程输出表、统计量和图形。

At last step, 只输出最终方程的表格、统计量和图形。

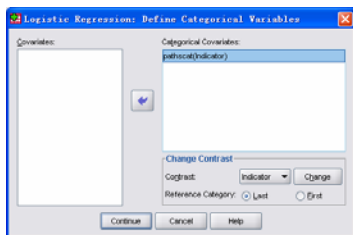


图 11-20 定义分类变量对话框

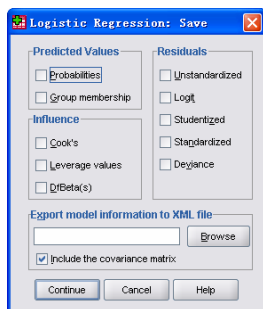


图 11-21 保存新变量对话框

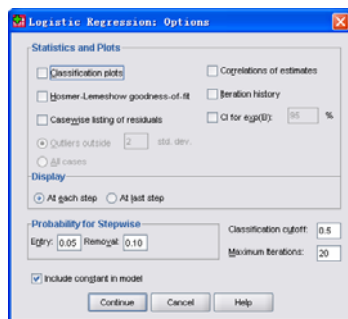


图 11-22 选项对话框

(3) Probability for Stepwise 栏, 设置变量进入模型及从模型中剔除的判据。如果变量的概率值小于 Entry 处的设置值, 那么此变量进入模型中, 如果其概率值大于 Removal 处的设置值, 变量会被从方程式中剔除。Entry 的默认值为 0.05, Removal 的默认值为 0.10。此处的设置值必须为正数, 而且 Entry 值必须小于 Removal 值。

(4) Classification cutoff 框, 设置系统划分观测量类别的辨别值。大于设置值的观测量被归于一组中, 反之观测量将被归于另一组中。其值的范围为 0.01~0.99。默认 0.5。

(5) Maximum Iterations 框, 输出最大的迭代步数。

(6) Include constant in model, 设置模型包括常数项。

### 11.3.3 二项逻辑斯谛回归分析实例

【例 4】data11-02 中是乳腺癌症患者的数据。利用 age 年龄、pathscat 扩散等级、pathsize 肿瘤尺寸变量, 建立一个预测因变量 ln\_yesno 癌变部位的淋巴结是否含有癌细胞的模型。

#### 1. 操作步骤

(1) 按 Analyze→Regression→Binary Logistic 顺序打开相应的对话框。

(2) 将变量 ln\_yesno 选入 Dependent 框, 将变量 pathsize、age、pathscat 作为自变量依次选入 Covariates 框。

(3) 打开 Categorical 对话框, 将变量 pathscat 选入 Categorical Covariates 框中, 在对比框中选择 Indicator 方式。对扩散等级变量 pathcat 重新编码为指示变量。

(4) 打开 Options 对话框, 选择 Classification Plots、Hosmer-Lemeshow goodness-of-fit 和 CI for exp(B), 在 Display 栏中选择 At last step 项。

(5) 在 SAVE 对话框中选择 Probabilities、Group membership 以便观察哪些患者属于淋巴结有癌细胞可能性较大。选择 Leverage value 通过杠杆值查找影响点。选择 Standardized 观察标准化残差以便使用图形对模型进行诊断。

其他选项为 SPSS 的默认选项, 单击 OK 按钮提交运算。

2. 输出结果见表 11-17~表 11-25 以及图 11-23。



表 11-17 为在计算过程中的观测量数量和缺失值的数量，以及它们所占的百分比。

表 11-18(a)为因变量变量的编码。(b)是自变量中的分类变量在模型中根据指示变量编码方案所生成的新变量表。新生成的变量名称为 pathscat(1)与 pathscat(2)。

表 11-17 观测量简表

Case Processing Summary			
Unweighted Cases <sup>a</sup>		N	Percent
Selected Cases	Included in Analysis	1121	92.9
	Missing Cases	86	7.1
	Total	1207	100.0
Unselected Cases		0	.0
Total		1207	100.0

a. If weight is in effect, see classification table for the total number of cases.

表 11-18 因变量与分类变量代码表

Dependent Variable Encoding		Categorical Variables Codings		
Original Value	Internal Value	Frequency	Parameter coding	
			(1)	(2)
无	0	肿瘤扩散等级 <= 2厘米	826	1.000
		2-5 厘米	293	.000
		> 5 厘米	12	.000
有	1			

(a)

(b)

表 11-19 起始模型外的变量

Variables not in the Equation					
Step 0	Variables	pathsize	Score	df	Sig.
		age	49.161	1	.000
		pathscat	31.793	1	.000
		pathscat(1)	34.262	2	.000
		pathscat(2)	26.897	1	.000
		pathscat(2)	19.449	1	.000
Overall Statistics			67.558	4	.000

表 11-20 第一步（最终）卡方检验表

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	64.897	4	.000
	Block	64.897	4	.000
	Model	64.897	4	.000

表 11-21 最终模型的拟合优度检验

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1151.770 <sup>a</sup>	.056	.085

表 11-22 Hosmer-Lemeshow 检验表

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	8.545	8	.382

表 11-23 Hosmer-Lemeshow 检验的列联表

Contingency Table for Hosmer and Lemeshow Test					
		癌变部位的淋巴结是否有癌细胞 是 = 无		癌变部位的淋巴结是否有癌细胞 是 = 有	
		Observed	Expected	Observed	Expected
Step 1	1	100	101.658	14	12.342
	2	102	95.735	9	15.265
	3	96	94.001	16	17.999
	4	88	90.962	23	20.038
	5	92	90.386	21	22.614
	6	86	87.109	26	24.891
	7	84	83.486	27	27.514
	8	74	81.573	39	31.427
	9	72	75.218	40	36.782
	10	66	59.871	46	52.129

表 11-24 最终观测量分类表

Classification Table <sup>a</sup>			
		Predicted	
		癌变部位的淋巴结是否有癌细胞 无	癌变部位的淋巴结是否有癌细胞 有
Step 1	癌变部位的淋巴结是否有癌细胞 无	846	14
	癌变部位的淋巴结是否有癌细胞 有	246	15
Overall Percentage			
			Percentage Correct
			98.4
			5.7
			76.8

a. The cut value is .500

表 11-25 最终模型统计量

Variables in the Equation								
		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)
Step 1 <sup>a</sup>	pathsize	.424	.131	10.487	1	.001	1.528	1.182 1.975
	age	-.025	.006	18.282	1	.000	.976	.965 .987
	pathscat			.548	2	.760		
	pathscat(1)	-.185	.846	.048	1	.827	.831	.158 4.362
	pathscat(2)	-.307	.728	.178	1	.673	.736	.176 3.066
	Constant	-.398	1.042	.146	1	.702	.671	

a. Variable(s) entered on step 1: pathsize, age, pathscat.

表 11-19 拟合起步前模型外的变量的卡方检验。所有单个变量 Sig 值均小于 0.05。4 个变量的总的卡方检验 Sig 值也小于 0.05。故均有资格进入模型。

表 11-20 是 3 种常用的卡方统计量。因为拟合方法选择的是默认的 Enter，只有一步

完成包含常数项与 5 个变量的模型的拟合, 所以模型的 Model、拟合过程块 Block 的和这一步 Step 的卡方值全部相同。如果采用的是逐步回归, 增加变量, 一步计算后 Sig 的值小于 0.05, 那么说明增加变量后的方程有意义; 剔除一个变量的一步后, 如果 Sig 的值大于 0.10, 那么说明剔除变量后的方程仍然有意义。

表 11-21 为模型拟合优度统计量。表中的  $-2ll$  值为 1151.770。此值较大, 说明模型对数据的拟合度不理想。接下来是 Cox & Snell  $R^2$  和 Nagelkerke  $R^2$  统计量, 其值分别为 0.056、0.085, 值太小, 说明能由方程解释的回归变异太少, 拟合效果不佳。

表 11-22 的 Hosmer-Lemeshow 是拟合统计量, 其零假设为方程对数据的拟合良好。本例 Sig>0.05, 无法拒绝零假设。这与表 11-23 的结论有差异, 故需要参考其他统计量。

表 11-23 以概率值为模型对淋巴结中是否含有肿瘤细胞进行 Hosmer-Lemeshow 检验的列联表。依据对观测量的预测 (淋巴结中是否含有肿瘤细胞) 概率, 它们被分为大致相等的 10 个组, Total 栏是每组观测量总数。由于将具有相同值的观测量组合在一起, 所以每组的观测量数并非精确地相等。在第 2、3 栏为观测到的和预测的淋巴结中不包含肿瘤细胞的数量, 4、5 栏为观测到的和预测的淋巴结中包含肿瘤细胞的数量。例如在第一个组中的 114 个观测量中实际有 14 个观测 (预测 12.3) 到淋巴结中包含肿瘤细胞, 100 个观测 (预测近 101.66) 到淋巴结中不包含肿瘤细胞, 其余各行的预测值与观测值都比较接近。

表 11-24 是以 0.5 作为淋巴结阳性与阴性 (淋巴结中包含、不包含肿瘤细胞) 分界线得出的预测值与实际数据的比较表。从表中可看到 846 名淋巴结中没有肿瘤细胞的观测对象被正确地预测, 正确率为 98.4%, 同时 246 名包含肿瘤细胞的患者被错误地预测为淋巴结中不包含恶性肿瘤细胞, 正确率仅为 5.7%。总的正确判断率为 76.8%。显然这个回归方程不能在实际中应用。据此可以估计淋巴结中发现癌细胞的概率

$$\text{prob}(\text{淋巴结中有癌细胞}) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

表 11-25 为模型中的各变量的相关统计量。根据表中各变量的系数 (B), 可以写出:

$$z = -0.398 + 0.424\text{pathsize} - 0.025\text{age} - 0.185\text{pathscat}(1) - 0.307\text{pathscat}(2)$$

图 11-23 为估计概率的分布图, 纵坐标为频数, 横坐标为预测淋巴结中含有癌细胞的概率值。图中的每个符号代表 5 个观测, 表示观测量预测归属的类别。如果模型拟合良好, 属于实际发生的观测量应该位于概率值 0.5 的右侧, 反之位于 0.5 的左侧。两组中的观测量越是分布在两端, 说明分组效果越好。图 11-23 中绝大部分观测量集中在 0~0.5 之间, 不同性质的观测量并没有正确地排列在图形的两端, 并且一些含有癌细胞的观测对象被错误地分在淋巴结中没有癌细胞的组中。显然把那些实际上淋巴结为阳性的对象预测为阴性要比把那些实际上淋巴结为阴性的对象预测为阳性所犯的误差更为严重。因此需要改变判别的标准, 以减少犯错误的可能性。例如可以设定只将那些估计概率小于 0.3 的观测对象称为“阴性”。

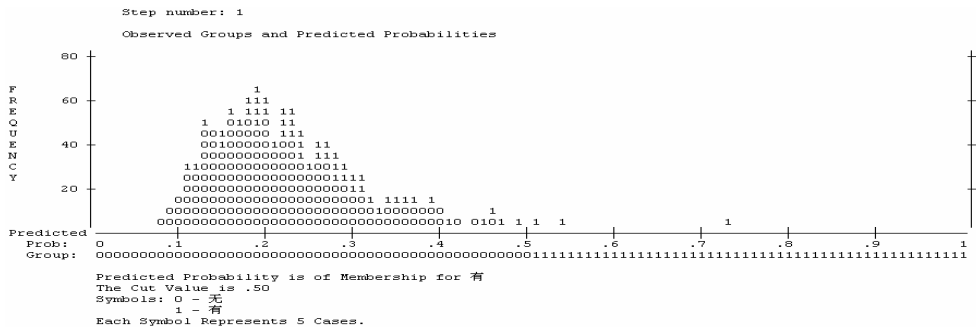


图 11-23 依据预测概率的观测量分组

如果改变分组的概率值后，大部分被错误分类的变量集中在概率值 0.3 左右，那么就应该停止使用设定的判别分组变量的判据。

3. 按 Graphs→Legacy Dialog→Scatter Dot 顺序单击菜单项，做杠杆值的散点图，见图 11-24。为预测概率与分类产生的新变量见图 11-25。

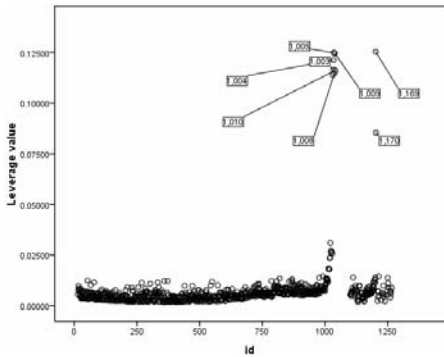


图 11-24 查找影响点的杠杆值图

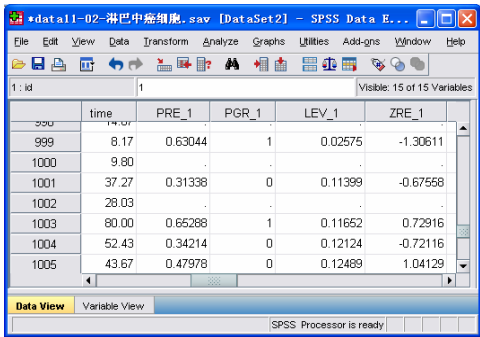


图 11-25 新变量：预测的概率与分类

Y 轴是新变量 LEV\_1 杠杆值，X 轴是 ID。杠杆值较大的对模型影响较大。双击该图进入图形编辑状态，对认为是影响点的离群点单击两次鼠标左键，在右键菜单中选择 Show data label 标出影响点的 ID 号。可以据此对这几个观测进行深入研究。

图 11-25 中的新变量。PRE\_1 是预测概率，PGR\_1 是预测分类。可以看到 PRE\_1 小于 0.5 的分到没有癌细胞的淋巴转移组，预测概率大于 0.5 的预测为有淋巴转移。

【例 5】计算一年龄为 60 岁，pathsize 肿瘤大小为 1 厘米，扩散等级 pathscat 为 2 的患者扩散到淋巴的概率。

注意，根据表 11-18 的编码方式，此例 pathscat(1)的值为 0，pathscat(2)的值为 1。

计算： $Z = -0.398 + 0.424 \times 1 - 0.025 \times 60 - 0.185 \times 0 - 0.307 \times 1 = -1.781$

其淋巴结中发现癌细胞的概率  $p = e^{-1.781} / (1 + e^{-1.781}) \approx 0.144 \approx 14.4\%$

在大多数情况下，如果此值小于 0.5，基本可以预测事件不会发生，大于 0.5 则反之。

结合查看图 11-23 也可以大致推测此人淋巴结中含有癌细胞的可能性不大。但由于模型可靠程度太低, 结论只能作为参考。

## 11.4 多分变量的逻辑斯谛回归

因变量为多水平分类变量的情况在医学领域中常见。比如在某一药物实验中, 动物服药后的状态是  $A$  (变量值为 1)、 $B$  (值为 2)、 $C$  (值为 3) 或是  $D$  (值为 4) 等。当因变量为多 (水平) 分类变量时, 可以使用多项逻辑斯谛回归的方法建立回归模型。

### 11.4.1 多分变量逻辑斯谛回归的概念

#### 1. 逻辑斯谛回归基本概念

对于因变量的  $k-1$  个水平, 每个水平一个回归方程, 每个水平的因变量概率值为 0~1。自变量是连续变量或计数变量 (非标称变量) 的, 可以用逻辑斯谛回归方法对因变量的概率值建立回归模型。回归曲线为典型的“S”形, 如图 11-16 所示。例如, 为了使得电影市场更加贴近观众, 可以使用电影观众的年龄、性别以及他们更喜欢观看的电影类型来建立多分变量逻辑斯谛回归, 预测常看电影的观众更喜爱哪种类型的影片。

Logistic 模型写为

$$\log\left(\frac{P(\text{event})}{1-P(\text{event})}\right) = b_0 + b_1x_1 + b_2x_2 + \Lambda + b_px_p$$

其中  $b_0$  为常数项,  $b_1$  到  $b_p$  为 Logistic 模型的回归系数, 是 Logistic 回归的估计参数,  $x_1$  到  $x_p$  为自变量。模型的左侧称为 Logit, 是事件发生几率的自然对数值。

如果因变量具有  $j$  类可能性, 第  $i$  类的模型为

$$\log\left(\frac{P(\text{category}_i)}{1-P(\text{category}_j)}\right) = b_{i0} + b_{i1}x_1 + b_{i2}x_2 + \Lambda + b_{ip}x_p$$

这样, 对于每一个 Logit 模型都将获得一组系数。例如, 如果因变量具有三种分类, 将会获得两组非零参数。

Logistic 回归方程另一种形式:  $P = \exp(y) / [1 + \exp(y)]$

其中  $y = a + \sum b_ix_i$  或  $y = \ln[P/(1-P)]$ , 通过变换可以得出  $P$  与变量  $x_i$  之间的数学表达式

$$P = \frac{\exp[a + \sum b_ix_i]}{1 + \exp[a + \sum b_ix_i]}$$

#### 2. 数据要求

因变量应该是分类变量, 自变量为因素变量与协变量 (因素变量必须为分类变量, 协变量必须是连续变量)。

#### 3. 模型检验

##### (1) 拟合检验

① Pearson 皮尔逊卡方统计量在多维表中检测观测频数与预测频数间的差异。公式为

$$X^2 = \sum_{\text{所有单元格}} \frac{(\text{观测数量} - \text{期望数量})^2}{\text{预测数量}}$$

其值越大, 显著性概率越低, 模型拟合效果越不好。

② 卡方偏差是另一个检测模型拟合度的指标。如果模型对数据拟合得好, 对数似然的差值就小, 显著性水平值越大。大样本数据的卡方偏差与皮尔逊卡方的值相近。

## (2) 伪 $R^2$ 统计量

在逻辑斯谛回归模型中使用 Cox & Snell、Nagelkerke 和 Mc Fadden 统计量。前两个已在前面介绍了, 这里介绍 McFadden 统计量。其公式为

$$R^2_{\text{McFadden}} = \frac{l(0) - l(B)}{l(0)}$$

式中  $l(B)$  为模型中对数似然比的核,  $l(0)$  为仅包含截距的模型的对数似然比的核。

## (3) 观测量—控制量的“配对”研究

它是一种利用现有观测数据研究那些很难发生的事件或是数据难以收集的事件。

例如: 汽车销售公司为了分析购买奔驰汽车客户的特点, 一般不得不收集大量的客户信息来确保分析的有效性, 而利用观测量—控制量的“配对”研究就可以不必收集很多购买了奔驰汽车的客户信息。这里, 观测量为已经有的购买了奔驰汽车的客户信息, 控制量是那些没有购买奔驰汽车的客户信息。观测量和控制量通过它们之间共有的年龄和性别进行配对。

① 对于包含  $k$  对观测量和控制量的数据, “经历”某种事件的 Logit 模型可以写成

$$\text{Log}(P_i) = \alpha_k + \sum_{i=1}^p b_i x_i$$

式中  $\alpha_k$  为根据配对变量值得到的第  $k$  对变量的“风险”,  $x_1$  到  $x_p$  为未配对自变量的值,  $b_i$  为第  $i$  个配对自变量的逻辑斯谛回归系数,  $P_i$  是事件的几率。

## ② 创建“差异变量”

SPSS 分析过程可以对满足特殊要求的一对一的变量数据进行分析。在配对分析中, 观测样本的样本量必须和与其配对的控制样本的样本量相同, 并且差异变量必须是配对的观测量与控制量间的差异。如果配对数多于 1 个, 则差异是平均值间的差。

现有 56 对母亲的数据, 其中一半的数据具有婴儿出生时较低体重的特点, 另一半没有这样的特点, 它们之间根据年龄 (配对变量) 配对。

其中的变量包括 lwt (怀孕前的体重)、age (年龄)、race (种族, 1: 白种人, 2: 黑色人种, 3: 其他入种); smoke (怀孕期间是否吸烟, 1: 吸烟, 0: 不吸烟); ptd (以前是否分娩, 0: 没有, 1: 有过); 以及 ui (子宫过敏, 1: 是, 0: 否)。

表 11-26 包含了配对后各变量之间的“差异”。虽然, 看起来计算观测量和控制量之间的差异比较容易, 但是当使用的分类变量超过两类时就会产生一些困难。在 SPSS 的逻辑斯谛回归过程中, 类似的分类变量必须事先定义为因素变量。在进行观测量—控制量配对分析时, 必须创建新变量替代分类变量, 并找到那些新变量之间的差异。

考虑一个简单的例子，种族变量具有 3 个值，所以需要使用两个变量表示它们。如果使用编码 1 表示参考类，必须创建两个新变量 `race1` 和 `race2`，其编码如表 11-27 所示。计算变量 `race1` 和 `race2` 之间的差值作为其他类。

### ③ 数据文件格式

表 11-26 配对变量之间的差异

	low	lwt	age	race	smoke	ptd	ui	race1	race2
观测量	1	101	14	3	1	1	0	0	1
控制量	0	135	14	1	0	0	0	0	0
差 异	1	- 34		X	1	1	0	0	1

表 11-27 种族的编码方式

	race1	race2
White	0	0
Black	1	0
Other	0	1

观测量—控制量的“配对”研究数据文件变量安排为：因变量、配对变量、观测量 1、控制量 1、差异变量 1、观测量 2、控制量 2、差异变量 2、…、观测量  $n$ 、控制量  $n$ 、差异变量  $n$ 。本例数据文件中应建立如下变量 `low`（因变量、其值应全部为  $1-0=1$ ）、`age`（配对变量）、`caslwt`（观测量 `lwt`）、`conlwt`（控制量 `lwt`）、`diflwt`（`lwt` 的差异变量）、`cassmoke`（观测量 `smoke`）、`cons smoke`（控制量 `smoke`）、`difsmoke`（`smoke` 的差异变量）…

最终将变量 `low` 设置为因变量，差异变量设置为协变量。

注意：如果交互项存在，首先必须创建交互项，然后计算它们之间的差异。在数据文件中的每一个观测量应该包含因变量、配对变量、控制变量、差异变量，以便将它们应用到相关的交互项分析中。对所有的观测量来说，因变量必须设置为一个常量，并且所有的差异变量必须设置为协变量（Covariates）。配对变量不能够作为主效应进入模型中，这是由于它们之间的差异为零。

## 11.4.2 多分变量的逻辑斯谛回归过程

1. 按 `Analyze→Regression→Multinomial Logistic` 顺序打开如图 11-26 所示的对话框。
2. 在左侧的源变量框中选择一个分类变量作为因变量送入 `Dependent` 框中。一般情况下 `Multinomial Logistic` 过程默认因变量的最后一类作为参考类，如果你要重新进行设置，单击 `Reference Category` 按钮进行设置，如图 11-27 所示。

(1) `Reference Category` 栏，设置参考类。`First` 或 `Last` 选项，第一类或最后一类作为参考类；`Custom` 选项后面由读者设置除第一和最后类以外的参考类。

(2) `Category Order` 栏。选择 `Ascending` 项，将分类变量中值最小的类设置为第一类，值最大的类设置为最后一类。选择 `Descending` 项，顺序相反。

3. 在源变量框中选择一个或多个分组变量送入 `Factor(s)` 框中。
4. 在源变量框中选择一个或多个连续变量作为协变量送入 `Covariates` 框中。
5. 单击 `Save` 按钮打开 `Save` 对话框，如图 11-28 所示。

(1) 在 `Save Variables` 栏选择要生成并保存到当前数据窗中的新变量：

- ① `Estimated response probabilities`，估计观测量进入因变量各组的响应概率值。
- ② `Predicted category`，预测的观测量分类。

③ Predicted category probability, 预测观测分类结果的概率。

④ Actual category probability, 实际分类的概率值。

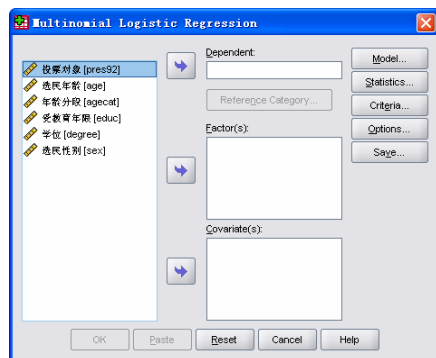


图 11-26 主对话框

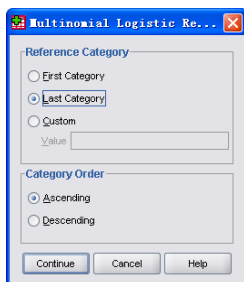


图 11-27 参考类对话框

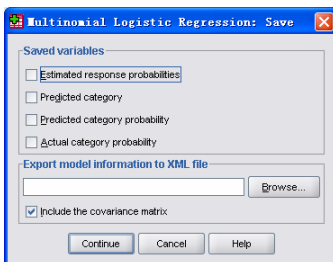


图 11-28 Save 对话框

(2) 在 Export model information to XMLfile 栏选择模型信息保存到外部 XML 格式文件中。

① 存储路径和文件名可以通过 Browse 按钮打开的对话框指定，也可以直接输入。

② Include the convarince matrix 要求输出的外部文件包括协方差矩阵。

6. 单击 Criteria 按钮打开判据对话框，见图 11-29，设置模型拟合过程结束的判据。

(1) 在 Iterations 栏，设置迭代停止的判据。

① Maximum Iterations 框，设置最大迭代数。必须为小于等于 100 的正整数。

② Maximum step-halving 框，输入使用 Step-halving 法的最大步数。

③ Log-likelihood convergence 框，设置对数似然比收敛值，必须为正数。回归过程中的对数似然比大于此值时，迭代过程停止。

④ Parameters convergence 框，设置收敛参数。如果在模型拟合过程中，绝对变化值或相对变化值大于等于此值时，迭代过程停止。

⑤ Print Iterations history for every \_\_ steps, 设置输出迭代过程的步距。

⑥ Check separation of data points from Iteration \_\_ forward 设置检查迭代过程开始值。

(2) Delta 框中输入小于 1 的非负值，此值会出现在交叉表的空单元中。这将有助于稳定算法、阻止估计偏差。Singularity tolerance 下拉列表中选择检验单一性的容许度值。

7. 主对话框中单击 Model 按钮打开模型对话框，如图 11-30 所示。Factors and Covariates 变量框中包含协变量和因素变量。

(1) Specifies model 栏中指定模型。

① Main effects, 主效应模型只包括协变量和因素变量的主效应。

② Full factorial model, 全模型中包含所有的主效应以及它们之间可能的交互效应。

③ Custom model, 自定义模型。模型中包括的主效应和交互效应都由读者指定。

(2) 以下选项只有在指定 Custom model 后生效。

① Build Teams 栏, 下拉列表中选择一种效应类型: main effects 主效应、interaction 交互效应、all 2-way 所有二维交互项, 以此类推。

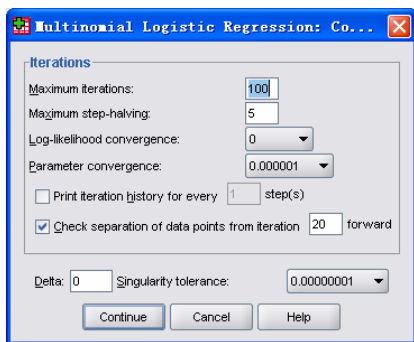


图 11-29 判据选择对话框

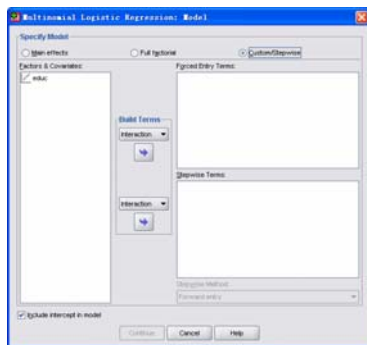


图 11-30 创建和选择模型对话框

② Forced Entry Terms 框, 选择强制出现在方程中的效应项进入此框。

③ Stepwise Terms 框, 选择要逐步加入或剔除出模型的效应项进入此框。

④ Stepwise method 下拉列表中可以选各效应项逐步进入方程的方法。有 Forward entry 向前进入、Backward elimination 向后剔除、Forward stepwise 逐步向前选择、Backward stepwise 逐步向后选择。

(3) Include intercept in model, 要求在模型中包含截距。

8. 主对话框中单击 Statistics 按钮打开如图 11-31 所示的对话框。选择输出统计量。

(1) Case processing summary, 给出分类变量综合信息。

(2) Model 栏选择模型统计量, 包括:

① Pseudo R-square, 输出 Cox & Snell、Nagelkerke  $R^2$  和 McFadden  $R^2$  统计量。

② Step summary, 在模型对话框选择了逐步筛选, 输出每一步变量进入或被剔除出方程时的效应表。

③ Model fitting information, 模型拟合信息。

④ Information Criteria 有关模型的判据信息。

⑤ Cell probabilities, 输出观测与期望频数表(带有残差)、协变量比率和响应分类。

⑥ Classification table, 输出每一类中观测和预测的分类表。

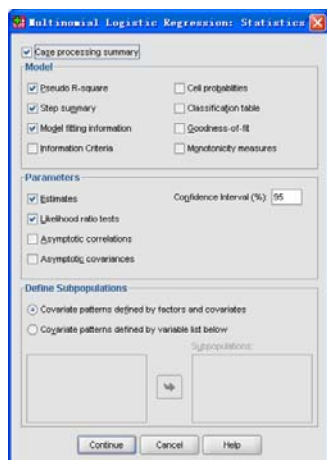


图 11-31 输出统计量对话框



⑦ Goodness-of-fit, 输出 Pearson 卡方和似然比卡方统计量。

⑧ Monotonicity measures 输出表中包括和谐对数、不和谐的对数和结点数, 和谐指数 C, 以及 Somers' D、Goodman、Kruskal's Gamma、Kendall's tau-a 等统计量。

(3) Paramete 栏, 指定要输出的与模型参数有关的统计量。

① Estimates, 模型的各种参数估计值, 包括由读者设置的置信区间。

② Likelihood ratio tests, 自动输出整个模型的检验统计量和模型的偏效应的似然比检验统计量。

③ Asymptotic correlations, 输出参数估计的相关阵。

④ Asymptotic covariances, 输出参数估计的协方差矩阵。

⑤ Confidence interval(%)框, 设置置信区间。

(4) Define Subpopulations 栏, 选择因素变量和协变量的子集以便定义协变量模式用于单元概率和拟合优度检验。

① Covariate pattern defined by factor and Covariate, 对所有因子变量和协变量进行拟合优度卡方检验。此为默认选项。

② Covariate pattern defined by variable list below, 在左下角的框中选择希望计算拟合优度卡方检验统计量的变量, 将其送入右下角的 Subpopulations 框中。

### 11.4.3 多分变量逻辑斯谛回归分析实例

【例 6】Data11-03 是 1992 年美国总统选举的数据, 用变量 sex 预测选民投票结果 pres92。

#### 1. 操作步骤

(1) 打开数据文件 Data11-03。按 Analyze→Regression→Multinomial Logistic 顺序打开相应对话框。

(2) 将投票变量 pres92 作为因变量选入 Dependent 框中; 将变量 sex 性别作为因素变量选入 Factor(s)框中; 在 Statistics 对话框中选择 Parameter 栏下的 estimate 复选项。

(3) 其他选项为 SPSS 的默认选项, 单击 OK 按钮提交运算。

2. 输出结果见表 11-32~表 11-35, Pres92 的值使用值标签。

表 11-28 为基本统计量包括: 投给布什、克林顿和帕洛特的票数。投票人性别比例。

表 11-29 是模型拟合信息, 最终方程的有效性检验, Sig 值小于 0.01, 因此方程有效。

表 11-30 为似然比统计量检测每一个变量对方程的影响, sex 变量的 Sig 值小于 0.01, 说明变量 sex 对方程具有重要意义。

表 11-31 Wald 统计量的 Sig 值全部小于 0.001, 因此可以将 Logit 模型写为

$$G1 = \log\left(\frac{P(\text{布什})}{P(\text{克林顿})}\right) = -0.5 + 0.433(\text{sex}) \quad G2 = \log\left(\frac{P(\text{帕洛特})}{P(\text{克林顿})}\right) = -1.51 + 0.715(\text{sex})$$

由于男性 sex 值为 1, 女性值为 0。因此简化了女性的 Logit 模型。例如, 第一个截距-0.5 解释为女性选布什的概率与选择克林顿概率之比的自然对数。

第二个截距 - 1.51 解释为女性选帕洛特的概率与选择克林顿概率之比的自然对数。变量 sex 的系数说明了 Logit 和性别之间的关系。因为所有的系数为正值并有显著意义。可以看出, 男性选布什和帕洛特的可能性要比女性大得多。

表 11-31 的系数描述了使用克林顿作为参照类别时不同性别的两个 Logit 模型, 同时获得了候选人之间的对比结果。也可以将布什与帕洛特进行对比, 根据  $\log(a/b)=\log(a)-\log(b)$ , 可以推出

$$\log\left(\frac{P(\text{布什})}{P(\text{帕洛特})}\right) = \log\left(\frac{P(\text{布什})}{P(\text{克林顿})}\right) - \log\left(\frac{P(\text{帕洛特})}{P(\text{克林顿})}\right)$$

查看参数 Exp(B) 可知男性选民选择布什的可能性是女性选民 1.54 倍 (布什与克林顿作比较), 选择帕洛特的可能性为女性的 2.04 倍 (帕洛特与克林顿作比较)。

3. 性别变量为什么能对投谁的票有很好的判断能力呢? 为分析不同性别对投票对象起决定作用是

否是因为不同性别的受教育的年限不同, 可以增加 educ 受教育程度作为协变量到模型中将变量 educ 送入 Covariates 栏, 其他操作同上, 运行结果见表 11-32。可以看出, 增加了 educ, 性别变量的系数改变很小。Wald 检验的零假设是回归系数均为 0。受教育年限的 Wald 统计量的 Sig 值全部大于 0.05, 说明无法拒绝该变量系数为 0 的假设。因此也可以试试将与受教育程度相关的学位变量作为因素来拟合模型。因为 educ 是连续变量, 而学位变量是分类变量, 可以作为因素变量。

表 11-28 基本统计量小结

Case Processing Summary		
	N	Marginal Percentage
对三个候选人的投票	布什	661 35.8%
	帕洛特	278 15.1%
	克林顿	908 49.2%
性别	男	804 43.5%
	女	1043 56.5%
Valid	1847	100.0%
Missing	0	
Total	1847	
Subpopulation	2	

表 11-29 模型拟合信息

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	61.209			
Final	27.343	33.866	2	.000

表 11-30 似然比卡方检验

Likelihood Ratio Tests				
Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	27.343 <sup>a</sup>	.000	0	
sex	61.209	33.866	2	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

表 11-31 模型参数估计

Parameter Estimates						
投票对象 <sup>a</sup>		B	Std. Error	Wald	df	Sig.
布什	Intercept	-.501	.068	54.067	1	.000
	[sex=1]	.433	.104	17.422	1	.000
	[sex=2]	0 <sup>b</sup>		0	0	
帕洛特	Intercept	-1.511	.098	235.703	1	.000
	[sex=1]	.715	.139	26.572	1	.000
	[sex=2]	0 <sup>b</sup>		0	0	

a. The reference category is: 克林顿

b. This parameter is set to zero because it is redundant.

4. 将性别变量和学位变量都作为因素变量做分析, 结果见表 11-32~表 11-35。

表 11-33 中的卡方值是排除因素变量与最终模型的两个 -2log Likelihood 的差值 Chi-Square。检验结果 Sig 小于 0.001, 说明最终模型成立。

表 11-34 是因素变量性别 sex、学历 degree 在最终模型中的似然比卡方检验结果。这是根据某个效应剔除出模型后的 -2ll 值的变化情况进行的检验, 其零假设为某变量被

从模型中剔除后该统计量没有变化。从表中的 Sig 值得出：sex 和 degree 剔除出模型后，-2ll 变化显著，拒绝性别和学历在模型中系数为 0 的假设。

表 11-35 是以克林顿为参考类模型中各参数及其检验结果。

表 11-32 变量 educ 作为协变量的模型参数及检验结果

Parameter Estimates							
投票对象 <sup>a</sup>		B	Std. Error	Wald	df	Sig.	Exp(B)
布什	Intercept	-.702	.259	7.318	1	.007	
	educ	.015	.018	.656	1	.418	1.015
	[sex=1]	.428	.104	16.970	1	.000	1.535
	[sex=2]	0 <sup>a</sup>			0		
帕洛特	Intercept	-1.894	.353	28.859	1	.000	
	educ	.027	.024	1.248	1	.264	1.028
	[sex=1]	.715	.139	26.396	1	.000	2.043
	[sex=2]	0 <sup>a</sup>			0		

a. The reference category is: 克林顿  
b. This parameter is set to zero because it is redundant.

表 11-33 模型拟合信息

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
	Intercept Only	178.457		
Final	103.601	74.856	10	.000

表 11-35 加入学位变量后参数估计及其检验

Parameter Estimates							
加入三个候选人的投票 <sup>a</sup>		B	Std. Error	Wald	df	Sig.	Exp(B)
布什	Intercept	-.805	.168	22.879	1	.000	
	[sex=1]	.458	.105	19.148	1	.000	1.581
	[sex=2]	0 <sup>a</sup>			0		
	[degree=0]	-.198	.228	.760	1	.383	.820
	[degree=1]	.387	.175	4.913	1	.027	1.473
	[degree=2]	.431	.253	2.914	1	.088	1.538
	[degree=3]	.424	.195	4.745	1	.029	1.529
	[degree=4]	0 <sup>a</sup>			0		
	帕洛特	Intercept	-2.188	.264	68.527	1	.000
	[sex=1]	.760	.140	29.319	1	.000	2.139
帕洛特	[sex=2]	0 <sup>a</sup>			0		
	[degree=0]	-.502	.393	1.627	1	.202	.605
	[degree=1]	.833	.267	9.709	1	.002	2.299
	[degree=2]	1.052	.346	9.263	1	.002	2.864
	[degree=3]	.804	.291	7.608	1	.006	2.233
	[degree=4]	0 <sup>a</sup>			0		

a. The reference category is: 克林顿  
b. This parameter is set to zero because it is redundant.

表 11-34 似然比卡方检验

Likelihood Ratio Tests				
Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	103.601 <sup>a</sup>	.000	0	.
sex	140.753	37.153	2	.000
degree	144.590	40.990	8	.000

以上结论没有考虑性别和学历之间的交互作用。下面进行这方面的研究。其他选项与前面相同，在 Model 选项中选择 Full factorial，在 Statistics 选项中选择 Likelihood ratio test，其输出结果如表 11-36 所示。

表11-36中，可以看到当交互项sex\*degree从方程式中剔除后，-2ll的变化值的Sig值很小，Sig大于0.05，也就是说“把它们剔除出模型时并没有改变模型的拟合程度”。因此采样表11-35中的参数进行进一步分析。

5. 计算预测概率和预期频数

根据逻辑斯蒂模型，可以计算一个选民投票给某个候选人的可能性。例如具有学士学位的男性选民投票给各候选人的可能性。

估计每个分类的概率的公式为

$$p(\text{group}_i) = \frac{\exp(g_i)}{\sum_{k=1}^j \exp(g_k)}$$

表 11-36 有交互项的似然比检验结果

Likelihood Ratio Tests				
Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	97.227 <sup>a</sup>	.000	0	.
sex	97.227 <sup>a</sup>	.000	0	.
degree	97.227 <sup>a</sup>	.000	0	.
sex * degree	103.601	6.374	8	.605

其中  $g_i$  是以最后一类做参考类, 第  $i$  类与参考类因变量值之比的概率的自然对数。本例很简单可以写出  $g1/g2$  的表达式计算其值, 根据三个候选人的被投票概率之和为 1, 列出联立方程得出解。

首先估算 3 个 Logit 模型的值, 根据表 11-35 的统计量可以分别计算出:

$$\begin{cases} \ln(p(\text{布什/克林顿})) = -0.805 + 0.458 + 0.424 = 0.077 \dots\dots\dots g1 \text{ 的值} \\ \ln(p(\text{帕洛特/克林顿})) = -2.188 + 0.760 + 0.804 = -0.624 \dots\dots\dots g2 \text{ 的值} \\ p(\text{布什}) + p(\text{帕洛特}) + p(\text{克林顿}) = 1 \end{cases}$$

解联立方程得到具有学士学位的男性选民, 对每一位候选人投票的可能性。

$$P(\text{布什}) = \frac{1.081}{(1 + 1.081 + 0.535)} = 0.413$$

$$P(\text{帕洛特}) = \frac{0.535}{1 + 1.081 + 0.535} = 0.205$$

$$P(\text{克林顿}) = \frac{1}{1 + 1.081 + 0.535} = 0.382$$

数据中有 160 名男性选民具有学士学位, 由此可以判断其中 66 人会投票给布什, 33 人会投给帕洛特, 61 人会投给克林顿。

表 11-37 观测值与预测值比较

6. 为了看到各类人实际投票与预测结果的比较, 可以在模型对话框中将 sex、degree 设置成主效应, 在 Statistics 对话框中选 Cell Probabilities 单元格概率, 选择 Classification table 分类表。运行结果参见表 11-37 和表 11-38。

表 11-37 是按性别、学历分组的实际和预测的单元格频数即百分比。表 11-38 模型实际预测的正确率的分类统计表, 在实际投给布什选票的 661 人中有 251 人, 大约占 38% 布什的支持者被模型正确地分类。没有一个帕洛特的支持者被正确地分类, 大约 3/4 的克林顿的支持者被模型正确地分类。总体来说, 被正确分类的约占近 50%。这说明模型对数据的分类效果不佳。当按因变量分组的观测在几组中的数量差别较大时, 无论模型拟合有多好, 根据统计量预测的结果, 总是会把更多的观测分入包含大数据量的组中。

RS最高学位	性别	对三个候选人的位置	Frequency			Percentage	
			Observed	Predicted	Pearson Residual	Observed	Predicted
低于高中	男	布什	27	27.902	-.210	32.5%	33.6%
		帕洛特	6	6.985	-.389	7.2%	8.4%
		克林顿	50	48.114	.419	60.2%	58.0%
	女	布什	28	27.098	.201	26.4%	25.6%
		帕洛特	6	5.015	.450	5.7%	4.7%
		克林顿	72	73.886	-.399	67.9%	69.7%
高中	男	布什	158	162.701	-.476	39.0%	40.2%
		帕洛特	89	86.103	.352	22.0%	21.3%
		克林顿	158	156.197	.184	39.0%	38.6%
	女	布什	191	186.299	.425	35.2%	34.4%
		帕洛特	70	72.897	-.365	12.9%	13.4%
		克林顿	281	282.803	-.155	51.8%	52.2%
大专	男	布什	22	21.965	.010	39.3%	39.2%
		帕洛特	17	13.856	.974	30.4%	24.7%
		克林顿	17	20.179	-.885	30.4%	36.0%
	女	布什	26	26.035	-.009	34.2%	34.3%
		帕洛特	9	12.144	-.984	11.8%	16.0%
		克林顿	41	37.821	.729	53.9%	49.8%
学士	男	布什	71	66.108	.785	44.4%	41.3%
		帕洛特	27	32.743	-1.125	16.9%	20.5%
		克林顿	62	61.149	.138	38.8%	38.2%
	女	布什	75	79.892	-.681	33.2%	35.4%
		帕洛特	35	29.257	1.138	15.5%	12.9%
		克林顿	116	116.851	-.113	51.3%	51.7%
研究生	男	布什	37	36.325	.140	37.0%	36.3%
		帕洛特	13	12.314	.209	13.0%	12.3%
		克林顿	50	51.361	-.272	50.0%	51.4%
	女	布什	26	26.675	-.155	28.0%	28.7%
		帕洛特	6	6.686	-.275	6.5%	7.2%
		克林顿	61	59.639	.294	65.6%	64.1%

The percentages are based on total observed frequencies in each subpopulation.

7. 为检验拟合的优劣，在 Statistics 对话框中选择 Goodness-of-fit，得出表 11-39。由于 Pearson、Deviance 统计量的 Sig 值全部大于 0.05，从而判断出模型对数据拟合较好。只有当协变量可以看做有序分类变量，且各分组单元都有大量的观测时，才能使用模型的拟合度统计量。如果协变量各单元格分组中观测量数差别很大，拟合度统计量不适用。

表 11-38 分类表

Classification				
Observed	Predicted			Percent Correct
	布什	帕洛特	克林顿	
布什	251	0	410	38.0%
帕洛特	133	0	145	0%
克林顿	237	0	671	73.9%
Overall Percentage	33.6%	0%	66.4%	49.9%

表 11-39 拟合度统计量

Goodness-of-Fit			
	Chi-Square	df	Sig.
Pearson	6.327	8	.611
Deviance	6.374	8	.605

11.5 概率单位回归

11.5.1 概率单位回归的概念

1. 概率单位回归分析

概率单位回归在 SPSS 软件中属于专业统计分析过程，用来分析反应比例与刺激强度之间的关系。例如研究一定数量的病人给药剂量与治愈的百分比之间的关系。

由于线性模型的某些限制，需要把可能分布在整个实数轴上的 x 值通过累计概率函数 f 变换成分布在(0,1)区间中的概率值，概率分布表达式为

$$P_i = f(\alpha + \beta x_i) = f(Z_i)$$

概率单位回归分析只考虑诸多累计概率函数中的两种：

(1) 标准正态累计概率函数：
$$P_i = F(Z_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-s^2/2} ds$$

式中  $P_i$  代表事件发生的概率，s 是零均值单位方差的正态分布的随机变量。由于这个概率是标准正态分布函数曲线下从  $-\infty$  到  $Z$  之间的面积，所以  $Z_i$  的值越大，事件就越可能发生。

(2) Logit 概率函数：
$$P_i = F(Z_i) = F(\alpha + \beta x_i) = \frac{1}{1 + e^{-Z_i}} = \frac{1}{e^{-(\alpha + \beta x_i)}} , \text{ 通过转换可以得到}$$

$$\text{Log} \frac{P_i}{1 - P_i} = Z_i = \alpha + \beta x_i$$

例如，可以设计一个实验，记录不同浓度的杀虫剂杀死的白蚁的数量。使用概率单位回归分析，就可以得出杀虫剂的浓度与杀死白蚁数量上的关系，据此判明什么样的杀虫剂浓度是最佳的（例如可以杀死 95% 以上的白蚁）。药学中，此方法常用于半数效量研究，即求完成 50% 反应的刺激量。

再如，可以用来检测购买某类物品的人员比例与所提供的物品刺激数量之间的关

系,在研究的数据具有相反的属性时(例如,买与不买),或者几组研究对象被作用于不同水平的刺激条件而产生不同的反应水平时才能应用概率单位回归分析。

## 2. 概率单位分析与 Logistic 分析的区别

概率单位模型实际上是由 Logit 模型和 Probit 模型组成。因此,首先利用 Logit 和(或)Probit 过程来转换响应比例,而不是直接使用“刺激”所产生的响应比例进行回归计算。

表 11-40 表明 Probit 和 Logit 的公式十分相似,因为 Logit 概率分布函数与正态分布密度函数近似,所以常用 Logit 模型来替代 Probit 模型。

## 3. 数据要求

- (1) 因变量中的每个数据应该是对某一水平刺激发生反应的数量。
- (2) 观测量应该是独立的,否则卡方检验和拟合优度检验是不适宜的。

表 11-40 概率分布函数数值比较

Z	正态累计概率函数 $p_i(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-s^2/2} ds$	Logit 概率函数 $p_i(Z) = \frac{1}{1 + e^{-Z_i}}$
-3.0	0.0013	0.0474
-2.0	0.0228	0.1192
-1.5	0.0668	0.1824
-1.0	0.1587	0.2689
-0.5	0.3085	0.3775
0.0	0.5000	0.5000
0.5	0.6915	0.6225
1.0	0.8413	0.7311
1.5	0.9332	0.8176
2.0	0.9772	0.8808
3.0	0.9987	0.9526

## 11.5.2 概率单位回归过程

1. 按 Analyze→Regression→Probit 顺序打开如图 11-32 的对话框。
  2. 选择一个变量作为响应频数变量进入 Response Frequency 框中。这个变量中的每一个数值是对实验刺激水平做出反应的观测量的数目总和,该变量的值不能为负数。
  3. 选择一个变量作为总观测变量进入 Total Observed 框中。这个变量是用于某一刺激水平的观测量总数。这个变量的值不能小于响应频数变量的值。
  4. 可选择一个因素变量进入 Factor 框。单击 Define 按钮,在对话框中给出因素变量的最小值和最大值。
  5. 选择至少一个协变量进入 Covariate(s)框中。协变量是不相同的实验刺激条件值。
- 协变量和 Probit(p)之间不存在线性关系时,在 Transform 框中选取转换模式,对协变量进行转换。三个选项分别为:None,不进行转换(默认);Log base 10,用以 10 为底的对数进行转换;Natural log,使用以 e 为底的自然对数进行转换。至于是否进行转换或选择哪种转换,要选择不同的转换方法,经过几次运行概率单位回归过程,比较分析结

果再确定。同时得出分析结论。

6. 在 Model 栏中确定一种算法

(1) Probit, 用累积标准正态分布函数的反函数来转换响应比例。

(2) Logit, 对响应比例应用自然对数转换。

7. 单击 Options 按钮, 打开如图 11-33 所示的对话框。

(1) Statistics 栏, 输出统计量。

① Frequencies, 输出每一个观测值与预测值的频数以及每一个观测值的残差。

② Relative median potency, 输出因素变量各水平间比较的效应及 95% 的置信区间。

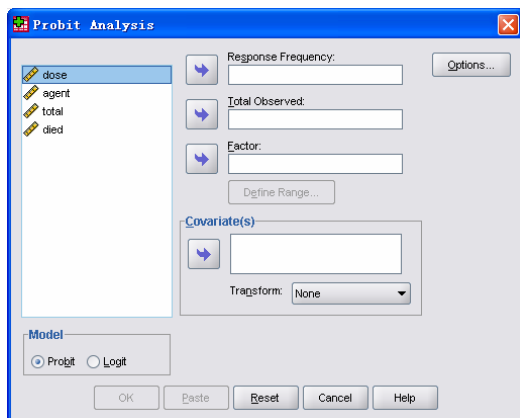


图 11-32 概率单位回归主对话框

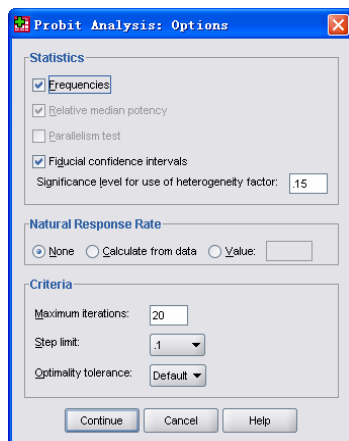


图 11-33 Options 对话框

③ Parallelism test, 平行检验的假设是因素变量各分组回归方程具有相同的斜率。

④ Fiducial confidence intervals, 如果选择了因素变量, 可选此项。在 Significance level for use of heterogeneity factor 框中输入一个显著性水平值。将对因素变量的每个水平显示从 0.01~0.99 反应比例所需的刺激强度的可信区间。当拟合优度值小于设定值时, Probit 用非齐性修正方法计算可信区间。选择了协变量, 就不适用置信区间与半数效量的计算。

(2) Natural Response Rate 栏, 设置是否计算自然响应率。在没有刺激条件下的响应称为自然响应。例如, 如果实验对象生命较短, 在实验过程中会发生一些自然死亡, 这时就需要调整观测比例以反映真实的“刺激”条件所产生的响应。

① None, 不计算自然响应率。

② Calculate from data, 根据提供的数据计算刺激强度为零的响应观测量。

③ Value, 输入小于 1 的已知自然响应频率。例如, 自然响应率是 12% 时, 输入 0.12。

(3) Criteria 栏, 设置控制迭代停止的判据。

① Maximum iterations 框, 输入控制迭代停止的最大迭代步数。

② Step limit 框, 选择参数向量所容许的最大变化量。

③ Optimality tolerance, 设定损失函数的精确值。

8. 单击 OK 按钮进行统计分析。

### 11.5.3 概率单位回归分析实例

【例 7】数据文件 data11-04 记录了不同杀虫药、不同浓度、不同杀虫效果的数据。变量包括: died 各组白蚁死亡数, total 各组白蚁总数, dose 杀虫剂剂量, agent 杀虫剂类别。使用这 4 个变量求各种杀虫剂的半数致死量。

#### 1. 操作步骤

(1) 按 Analyze→Regression→Probit 顺序打开对话框。

(2) 选择 died 作为响应变量送入 Response Frequency 框中; 选择 total 作为总观测量变量送入 Total observed 框中。选择剂量变量 dose 送入 Covariate(s)框中。

(3) 选择 agent 作为因素变量送入 Factor 框中, 单击 Define Range 按钮, 打开对话框, 在 Minimum 后输入 1, 在 Maximum 后输入 3。

(4) 在 Transform 下拉列表中选择 Log Base 10, 作为第一次运行该分析过程的选择。

(5) 在 Options 对话框选择 Parallelism test, 其他参数选项均为默认值。

(6) 单击 OK 按钮进行统计分析。结果输出见表 11-41~表 11-42 和图 11-33~图 11-36。

#### 2. 结果分析

表 11-41 数据基本统计

Data Information		N of Cases
Valid		15
Rejected	Out of Range <sup>a</sup>	0
	Missing	0
	LOG Transform Cannot be Done	0
	Number of Responses > Number of Subjects	0
Control Group		0
agent	鱼藤素	5
	鱼藤酮	5
	混合物	5

a. Cases rejected because of out of range group values.

表 11-42 模型参数

Convergence Information					
	Number of Iterations	Optimal Solution Found			
PROBIT	15	Yes			

Parameter Estimates						
Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
PROBIT <sup>a</sup>						
Intercept <sup>a</sup>	4.006	.274	14.640	.000	3.469	4.542
deguelin	-2.743	.214	-12.800	.000	-2.958	-2.529
rotenone	-4.492	.386	-12.274	.000	-4.858	-4.126
mixture	-2.741	.214	-12.809	.000	-2.955	-2.527

a. PROBIT model: PROBIT(g) = Intercept + BX (Covariates X are transformed using the base 10.000 logarithm.)

b. Corresponds to the grouping variable agent.

Chi-Square Tests			
	Chi-Square	df <sup>a</sup>	Sig.
PROBIT			
Pearson Goodness-of-Fit Test	9.374	11	.587 <sup>a</sup>
Parallelism Test	1.664	2	.435

a. Statistics based on individual cases differ from statistics based on aggregated cases.

b. Since the significance level is greater than .150, no heterogeneity factor is used in the calculation of confidence limits.

图 11-41 给出了数据的基本情况。共有 15 个合法观测量, 没有观测量被剔除, 三种杀虫剂鱼藤素 deguelin、鱼藤酮 rotenone、混合物 mixture 的观测量数均为 5 个。

在图 11-42 中, 第 1 个表说明进行 15 步迭代后, 找到了最佳结果;

第 2 个表是参数估计表。给出了方程形式, 三种白蚁杀虫剂效果的模型为:

杀虫剂 deguelin 鱼藤素的方程:  $Probit(p) = -2.743 + 4.006 \lg(dose)$ ;

杀虫剂 rotenone 鱼藤酮的方程:  $Probit(p) = -4.492 + 4.006 \lg(dose)$ ;



杀虫剂 mixture 混合物的方程： $Probit(p) = -2.741 + 4.006 \lg(dose)$ 。

第 3 个表，皮尔逊拟合优度卡方检验的显著水平 0.587 大于 0.05，拟合良好。

如果皮尔逊卡方显著水平值较小，或许是因为药剂量与  $Probit(p)$  之间没有存在线性关系，或虽为线性，但观测量在直线周围的分布不均匀。

由于 Parallelist test 平行检验的  $P$  值为 0.435 大于 0.05，不足以拒绝零假设（不排除在更多样本时，或另一个检验方法时拒绝零假设）。即三种杀虫剂方程式直线相互平行。

表 11-43 为三种杀虫剂各剂量 dose 的致死率 Prob 及 95%置信区间上下限。表中查三种杀虫剂的半数致死量，即  $Prob=0.5$  时的剂量的估计值分别为 4.840、13.229、4.833。

表 11-43 三种杀虫剂各剂量致死率与 95%的置信区间

Confidence Limits															
		95% Confidence Limits for dose													
agent	Probability	Estimate	Lower Bound	Upper Bound											
PROBIT	deguelin	0.01	1.271	998	1.538	rotenone	0.01	3.473	2.589	4.381	mixture	0.01	1.269	846	1.921
		0.02	1.487	1.192	1.772	0.02	4.063	3.094	5.045	0.02	1.484	795	2.181		
		0.03	1.642	1.335	1.938	0.03	4.487	3.463	5.519	0.03	1.640	906	2.365		
		0.04	1.769	1.452	2.074	0.04	4.836	3.768	5.906	0.04	1.767	1.000	2.513		
		0.05	1.880	1.555	2.192	0.05	5.139	4.036	6.241	0.05	1.878	1.083	2.641		
		0.06	1.980	1.649	2.298	0.06	5.412	4.279	6.542	0.06	1.977	1.159	2.755		
		0.07	2.072	1.735	2.395	0.07	5.664	4.504	6.818	0.07	2.069	1.231	2.859		
		0.08	2.158	1.816	2.486	0.08	5.899	4.714	7.075	0.08	2.155	1.298	2.956		
		0.09	2.240	1.893	2.572	0.09	6.121	4.914	7.318	0.09	2.236	1.363	3.047		
		0.1	2.317	1.966	2.654	0.1	6.333	5.106	7.549	0.1	2.314	1.425	3.133		
		0.15	2.668	2.299	3.023	0.15	7.291	5.977	8.592	0.15	2.664	1.714	3.518		
		0.2	2.984	2.601	3.357	0.2	8.155	6.770	9.529	0.2	2.980	1.983	3.858		
		0.25	3.295	2.890	3.676	0.25	8.977	7.528	10.420	0.25	3.280	2.247	4.179		
		0.3	3.581	3.171	3.992	0.3	9.786	8.277	11.298	0.3	3.576	2.513	4.490		
		0.35	3.879	3.454	4.312	0.35	10.600	9.033	12.184	0.35	3.873	2.787	4.802		
		0.4	4.184	3.743	4.645	0.4	11.436	9.808	13.095	0.4	4.178	3.072	5.120		
		0.45	4.503	4.041	4.995	0.45	12.307	10.616	14.049	0.45	4.497	3.375	5.450		
		0.5	4.840	4.355	5.370	0.5	13.229	11.469	15.065	0.5	4.833	3.700	5.799		
		0.55	5.203	4.688	5.779	0.55	14.219	12.383	16.165	0.55	5.195	4.053	6.175		
		0.6	5.599	5.040	6.232	0.6	15.302	13.377	17.376	0.6	5.591	4.443	6.587		
0.65	6.040	5.444	6.744	0.65	16.508	14.477	18.738	0.65	6.032	4.879	7.049				
0.7	6.543	5.888	7.337	0.7	17.882	15.720	20.306	0.7	6.534	5.378	7.583				
0.75	7.133	6.403	8.044	0.75	19.494	17.165	22.170	0.75	7.123	5.962	8.222				
0.8	7.852	7.020	8.923	0.8	21.459	18.906	24.478	0.8	7.841	6.665	9.025				
0.85	8.783	7.803	10.082	0.85	24.002	21.125	27.517	0.85	8.770	7.554	10.110				
0.9	10.111	8.960	11.778	0.9	27.634	24.235	31.958	0.9	10.097	8.768	11.761				
0.91	10.461	9.185	12.232	0.91	28.591	25.044	33.146	0.91	10.446	9.076	12.216				
0.92	10.855	9.504	12.745	0.92	29.667	25.949	34.491	0.92	10.840	9.418	12.737				
0.93	11.306	9.987	13.337	0.93	30.998	26.977	36.040	0.93	11.290	9.802	13.345				
0.94	11.831	10.287	14.031	0.94	32.333	28.169	37.859	0.94	11.814	10.242	14.069				
0.95	12.440	10.787	14.869	0.95	34.052	29.595	40.055	0.95	12.442	10.758	14.957				
0.96	13.241	11.403	15.920	0.96	36.188	31.331	42.812	0.96	13.222	11.385	16.090				
0.97	14.269	12.207	17.319	0.97	38.998	33.606	46.481	0.97	14.249	12.188	17.627				
0.98	15.761	13.359	19.377	0.98	43.074	36.865	51.880	0.98	15.738	13.318	19.938				
0.99	18.435	15.390	23.142	0.99	50.381	42.604	61.766	0.99	18.408	15.262	24.295				

表 11-44 是按因素变量分组所得的观测值与期望值数据。杀虫剂类别 agent 为分组变量，dose 为剂量，Number of Subjects 为观测量总数，Observed Responses 为响应频数观测值，Expected Responses 为响应频数期望值，Residual 为残差，Probability 为概率。

表 11-45 为各组半数效应比值，杀虫剂 deguelin 对 rotenone 比值为 4.84/13.22=0.366，mixture 对 rotenone 比值为 1.001，rotenone 对 mixture 比值为 2.737。

图 11-34 为三种杀虫剂剂量取对数与概率值的散点图，从图中可以看出概率值与不同刺激剂量呈现较为明显的线性关系，说明取“Log Base 10”的选项进行转换是比较合适的，如果散点图没有呈现线性关系，那么还需要进行其他方法的转换，或各种转换各做一次，比较其结果。一定要确保转换后数据的线性关系。

表 11-44 观测与期望频数

Cell Counts and Residuals								
	Num ber	agent	药物剂量	Number of Subjects	Observed Responses	Expected Responses	Residual	Probability
PROBIT	1	1	.410	50	6	6.769	-.769	.135
	2	1	.580	48	16	16.170	-.170	.337
	3	1	.710	46	24	24.852	-.852	.540
	4	1	.890	49	42	38.916	3.084	.794
	5	1	1.010	50	44	45.176	-1.176	.904
	6	2	1.000	48	18	15.035	2.965	.313
	7	2	1.310	48	34	37.198	-3.198	.775
	8	2	1.480	49	47	45.301	1.699	.925
	9	2	1.610	50	47	48.741	-1.741	.975
	10	2	1.700	48	48	47.508	.492	.990
	11	3	1.009	50	44	45.153	-1.153	.903
	12	3	.886	49	42	38.762	3.238	.791
	13	3	.708	46	24	24.712	-.712	.537
	14	3	.580	48	16	16.214	-.214	.338
	15	3	.415	50	6	7.019	-1.019	.140

表 11-45 各组半数效应比较值

组别		95% Confidence Limits			95% Confidence Limits with LOG Transform*		
		Estimate	Lower Bound	Upper Bound	Estimate	Lower Bound	Upper Bound
PROBIT	1	.366	.246	.500	-.437	-.609	-.301
	2	1.001	.884	1.161	.001	-.063	.065
	3	2.733	1.998	4.066	.437	.301	.609
	4	2.737	2.000	4.074	.437	.301	.610
	5	.365	.245	.500	-.437	-.610	-.301
	6	.999	.881	1.157	.000	-.065	.063

a. Logarithm base = 10.

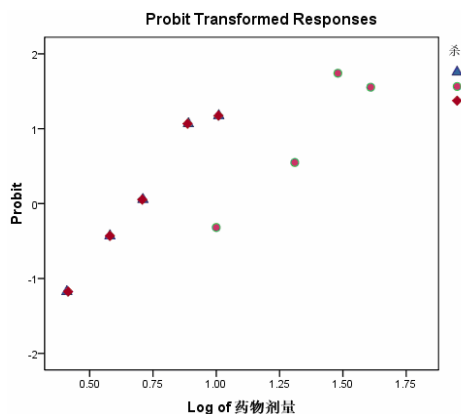


图 11-34 散点图

## 11.6 非线性回归

### 11.6.1 非线性模型

#### 1. 本质线性模型与本质非线性模型

$$y = e^{b_0 + b_1 x_1 + b_2 x_2 + e},$$

上式所表达的模型只要两边取自然对数，就可以写为

$$\ln(y) = b_0 + b_1 x_1 + b_2 x_2 + e$$

这种看起来非线性，但可以转换为线性的模型，称为本质线性模型。

当把一个模型转换为线性模型后，必须确保转换后的误差项也要满足所需的假设条件。例如，对于原始方程  $y = e^{bx} + e$ ，由于取对数后失去误差项  $e$ ，为了保证在转换后的模型中也存在误差项，原始的方程式应该写为

$$y=e^{bx+e}=e^{bx}e^e$$

$$y=B_0+e^{b_1x_1}+e^{b_2x_2}+e^{b_3x_3}+e$$

不能转换为线性模型，称为本质非线性模型。在非线性回归过程中，必须首先估算将会应用到非线性模型中的起始值和参数值的范围，目的只是要将残差平方和减少到最小。本节解决本质非线性问题。

2. 常用非线性模型

表 11-46 是已经得到公认的，并且经常使用的非线性模型。

注意，不能随意套用。

表 11-46 常用非线性模型

名 称	模 型 表 达 式
Asymptotic	$b_1 + b_2 * \exp(b_3 * x)$
Asymptotic	$b_1 - (b_2 * b_3^x)$
Density	$(b_1 + b_2 * x)^{(-1/b_3)}$
Gauss	$b_1 * (1 - b_3 * \exp(-b_2 * x^2))$
Gompertz	$b_1 * \exp(-b_2 * \exp(-b_3 * x))$
Johnson-Schumacher	$b_1 * \exp(-b_2 / (x + b_3))$
Log-Modified	$(b_1 + b_3 * x)^{b_2}$
Log-Logistic	$b_1 - \ln(1 + b_2 * \exp(-b_3 * x))$
Metcherlich Law of Diminishing Returns	$b_1 + b_2 * \exp(-b_3 * x)$
Michaelis Menten	$b_1 * x / (x + b_2)$
Morgan-Mercer-Florin	$(b_1 * b_2 + b_3 * x^{b_4}) / (b_2 + x^{b_4})$
Peal-Reed	$b_1 / (1 + b_2 * \exp(-(b_3 * x + b_4 * x^2 + b_5 * x^3)))$
Ratio of Cubics	$(b_1 + b_2 * x + b_3 * x^2 + b_4 * x^3) / (b_5 * x^3)$
Ratio of Quadratics	$(b_1 + b_2 * x + b_3 * x^2) / (b_4 * x^2)$
Richards	$b_1 / ((1 + b_3 * \exp(-b_2 * x))^{(1/b_4)})$
Verhulst	$b_1 / (1 + b_3 * \exp(-b_2 * x))$
Von Bertalanffy	$(b_1^{(1-b_4)} - b_2 * \exp(-b_3 * x))^{(1/(1-b_4))}$
Weibull	$b_1 - b_2 * \exp(-b_3 * x^{b_4})$
Yield Density	$(b_1 + b_2 * x + b_3 * x^2)^{-1}$

3. 条件逻辑表达式

条件逻辑表达式应用于方程中或损失函数（Loss function）中。为了表达一个模型中或损失函数中的条件逻辑式，必须将几个不同条件的分段模型组合在一起。每一个分段模型由逻辑表达式乘以逻辑表达式为真时的结果。例如，分段模型表示为

$$\hat{f}(x) = \begin{cases} 0 & x \geq 0 \\ x & 0 < x < 1 \\ 1 & x \leq 1 \end{cases}$$

这几个分段模型组合后的逻辑表达式为 $(x \leq 0) * 0 + (x > 0 \ \& \ x < 1) * x + (x \geq 1) * 1$ ，因为逻辑表达式的值只能是 1（真）或 0（假），因此

如果  $x \leq 0$ , 以上结果为  $1*0 + 0*x + 0*1 = 0$ ;

如果  $0 < x < 1$ , 以上结果为  $0*0 + 1*x + 0*1 = x$ ;

如果  $x \geq 1$ , 以上结果为  $0*0 + 0*x + 1*1 = 1$

两个不等式之间必须由逻辑运算符连接。例如:  $0 < x < 1$  必须写成  $(x > 0 \ \& \ x < 1)$ 。

字符串表达式可以被用于逻辑表达式中。(sex='M')\*worth + (sex='F')\*0.59\*worth 的结果为: 当变量 sex 值为 M 时变量 worth 的值, 与变量 sex 值为 F 时变量 worth 的值乘 59%之和。

#### 4. 损失函数

在非线性回归中, 损失函数是对某统计量的运算法则, 非线性回归过程以将其值最小化为原则进行非线性拟合。SPSS 默认根据最小残差平方和找出非线性模型。也可以自定义损失函数。

#### 5. 参数约束

在多数的非线性模型中, 参数必须限制在有意义的区间中。所谓约束是指在利用迭代方法求解的过程中对参数值的限制。可以首先使用线性约束, 防止结果溢出。

(1) 线性约束: 将参数乘以常数, 该常数不能是其他参数或者自身。

(2) 非线性约束: 其中至少一个参数与其他参数相乘或相除或者进行幂运算。

#### 6. 数据要求

因变量和自变量应该是数值型变量。名义变量应该被重新编码为二分(哑)变量或者是其他类型的对比变量。同时要求定义的函数要尽可能精确地反映因变量与自变量之间的关系。

#### 7. 估算初始值

即使模型是非常精确的, 准确地确定参数的初始值也是非常重要的。为参数设置合适的初始值以保证正常、迅速收敛, 同时避免解决方案范围小于实际范围。

(1) 使用图形辅助确定参数取值范围, 在研究的实际范围内确定初始值。

(2) 根据确定的非线性方程的数学特性进行变换, 结合图形辅助判断初始值范围。

(3) 直接使用数值来替代某些参数, 确定其他参数的取值范围, 从而确定初始值。

(4) 将数据转换后, 使用线性关系模型确定初始值。

通常联合使用上述几种方法。如果参数没有初始值, 也不要仅仅将它们设置为 0, 最好是将它们设置为预计要改变的值的的大小。如果忽略误差项, 或许可以获得一个线性模型, 并根据线性模型估算初始值。例如方程式  $y = e^{a+bx} + \varepsilon$ , 如果忽视误差项  $\varepsilon$ , 并且在两边取对数, 获得方程式  $\ln(y) = a + bx$ , 就可以利用线性模型来估计参数  $a$ 、 $b$  的值了。

#### (5) 利用非线性模型的属性估算初始值

有时能确定因变量在一定范围内的值。例如, 如果在模型  $y = e^{a+bx}$  中, 当  $x=0$  时,  $y=2$ , 就可以取  $\ln 2$  的值作为参数  $a$  的初始值。考虑当方程的值为最大值和最小值, 或者当所有的自变量接近 0 时, 或其值接近无限大时的情况, 会对确定参数的起始值有帮助。

(6) 利用与参数同等数量的方程式，可以解决参数的初始值问题。再看前面的例子，可以解联立方程

$$\ln(y_1)=a+bx_1$$

$$\ln(y_2)=a+bx_2$$

利用减法得  $\ln(y_1)-\ln(y_2)=bx_1-bx_2$

解此方程式，得参数  $b=\frac{\ln(y_1)-\ln(y_2)}{x_1-x_2}$ ， $a=\ln(y_1)-bx_1$

11.6.2 非线性回归过程

- 1. 按 Analyze→Regression→Nonlinear 顺序打开如图 11-35 所示的非线性回归主对话框。从源变量框中选择一个数值型变量作为因变量送入 Dependent 框中。
- 2. 在 Model Expression 框中输入合适的模型表达式，其中应至少包括一个自变量。
  - (1) 将变量选入 Model Expression 框中。字符型变量仅能在逻辑表达式中使用。
  - (2) 定义模型表达式。

从 Function group 框中选择需要的非线性函数送入 Model Expression 模型表达式框中。从计算模板上选择数字或操作符，送入模型表达式的相应位置，组成模型表达式。注意，参数名不能与所选择的变量同名。

- (3) 定义模型参数。

单击 Parameters 按钮，打开如图 11-36 所示的 Parameters 对话框。在 Name 框中输入参数名。在 Starting Value 框中输入尽量准确的初始值，即尽可能接近期望值。定义一个单击 Add 按钮确定一个。直到把所有参数定义完，选择某个参数，单击 Remove 按钮可将其剔除；修改后单击 Change 按钮确认。定义或修改完成，单击 Continue 按钮回主对话框。这里设置的参数以及初始值将在以后的分析中一直起作用。

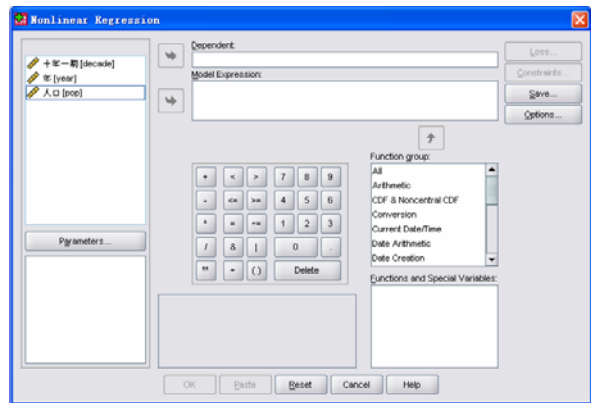


图 11-35 非线性回归主对话框

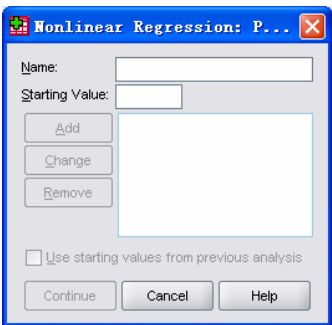


图 11-36 参数设置

如果前次运行非线性函数，参数显示在 Parameters 框中。要使用这些参数做初始值。在参数对话框中选择 Use starting value from previous analysis。改变了模型表达式不能选择此项。

3. 如果需要对 Parameters 框中的参数取值范围进行约束，单击 Constraints 按钮，打开如图 11-37 所示的对话框。

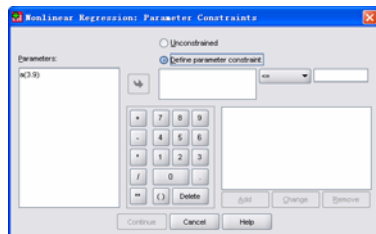


图 11-37 参数约束对话框

(1) Unconstrained，默认对参数的值不限制。

(2) Define parameter constraint，定义对参数的限制。在参数框中选择需要约束的参数送入 Define parameter constraint 框中。在逻辑运算符下拉列表中  $\leq$ 、 $\geq$ 、 $=$  选择 3 中的一个，在最右侧的框中输入常数。构成约束表达式，单击 Add 按钮送入右下角的框中。选择表达式单击 Remove 按钮可将其删除；修改后单击 Change 按钮确认，显示新表达式。选择 Continue 按钮返回主对话框。

4. 在非线性回归中默认的损失函数残是差平方和。要自定义损失函数，在参数框中选择一个或多个参数，然后单击 Loss 按钮，打开如图 11-38 所示的损失函数对话框。

(1) Sum of squared residuals 残差平方和是系统默认的损失函数。

(2) 选择 User-defined loss function，使用定义的损失函数。选择此项，输入自定义的损失函数。RESID\_表示残差，PRED\_表示预测值。规定 RESID\_2 表示残差的平方和。

5. 单击 Options 按钮，打开 Options 对话框，见图 11-39，在对话框中确定标准误的估计方法或者确定迭代过程停止的判据。

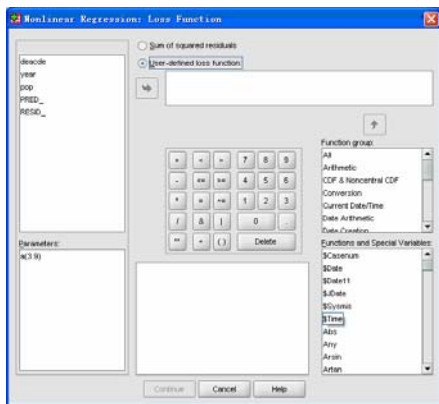


图 11-38 损失函数对话框

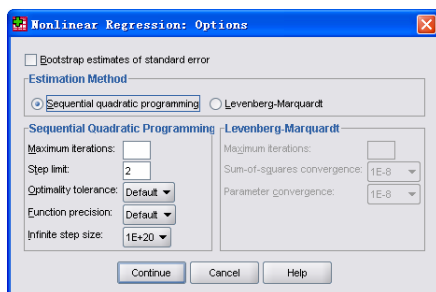


图 11-39 选择对话框

(1) Bootstrap estimates of standard error，标准误的自举估计是反复从原始数据集中提取相同数量的样品计算标准误的一种方法。针对每一个样本建立相应的非线性回归模型，计算每个参数估计的标准误作为自举估计的标准差。原始数据的参数值作为每一个自举

样本的初始值。

(2) Estimation Method 栏, 选择估计方法。

① Sequential quadratic programming, 适用于限制模型与非限制模型。如果确定了一个限制模型、定义了损失函数或选了自举估计, 则自动选中该项。它利用双重迭代算法求解, 每一步迭代建立一个二次规划算法, 确定寻找的方向, 并在选择的方向中发现一个新点, 而损失函数对新点进行求值, 直到寻找过程发生收敛。判据和精度选项如下:

- Maximum iterations 框中输入最大迭代步数作为迭代停止的判据。
- Step limit 框, 输入一个正值作为参数向量长度的最大允许变化量。
- Optimality tolerance, 在下拉列表中选择最优容限, 即目标函数的精度, 即有效位数。如果容限为 0.1E-6, 有效数字为六位。最优容限值必须大于函数精度。
- Function precision 下拉列表中选择小于最优容限并在 0~1 之间的数字, 作为目标函数精度。函数值较大时, 作为相对精度; 函数值较小时, 看做是绝对精度。
- Infinite step size, 在一步迭代过程中参数的变化大于设置值, 迭代停止。

② Levenberg-Marquardt, 非线性约束模型的默认运算法则, 如果确定了一个线性约束模型, 或者定义了一个损失函数, 或者选中标准误的自举估计, 那么该选项不起作用。控制迭代停止的判据有:

- Maximum iterations, 输入 Levenberg-Marquardt 算法中最大的迭代步数。
- Sum of squares convergence 框, 残差平方和的变化量小于设置值, 迭代停止。
- Parameter convergence, 任何一个参数值的变化小于设置值, 迭代停止。

后两项的默认值均为 1E-8。

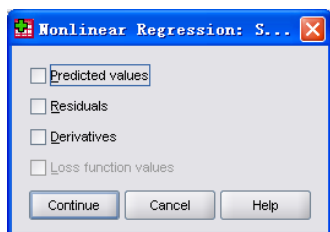


图 11-40 保存新变量对话框

6. 单击 Save 按钮, 打开如图 11-40 所示的对话框。指定要生成的新变量:

- (1) Predicted values, 因变量预测值, 变量名为 Pred\_。
- (2) Residuals 残差, 变量名为 Resid。
- (3) Derivatives 派生项, 变量名为参数名后加字母“d”。
- (4) Loss function values, 损失函数值。定义了损失函数, 才会保存损失函数变量值, 其变量名称为 Loss\_。

### 11.6.3 非线性回归分析实例

【例 8】Data11-05 是美国 1790~1960 年人口变化的数据, 人口单位为百万。图 11-41 为人口与年代的散点图。根据经验, 对于人口数量模型的建立经常使用 Logistic 模型, 其方程为

$$y_i = \frac{c}{1 + e^{a+bt_i}} + e_i$$

其中  $y_i$  是在时间  $t_i$  时的人口数量,  $e_i$  为误差项,  $a$ 、 $b$  为参数。虽然通常模型对观测数据的拟合程度相当好, 但有关误差项的独立性假设和常数项方差的假设却有可能被破坏。这是由于时间序列的数据误差项往往并不独立, 误差项的大小有可能依数据总体的大小而变化。由于人口成长的模型不能被转换为线性模型, 因此选择非线性模型来估算模型的参数。

### 1. 初始值的确定

本例利用简单的假设来确定初始值。在 Logisitic 人口增长模型中, 参数  $c$  为渐近线。任意选择距最大观测值不远的渐近线。本例最大人口值为 178, 故选择 200 做  $c$  的初始值, 然后依据时间为 0 的人口值来估算参数  $a$  的值

$$3.895 = \frac{200}{1 + e^{a+b*0}}$$

$$a = \ln\left(\frac{200}{3.895} - 1\right) = 3.9$$

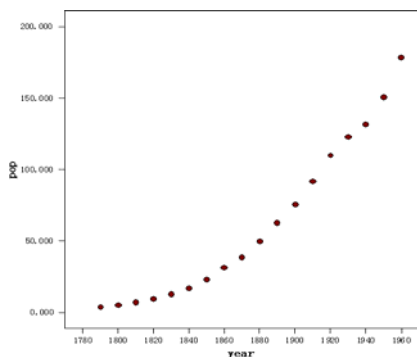


图 11-41 美国 1790—1960 年人口散点图

接下来利用时间为 1 时的人口值来估算参数  $b$  的初始值, 如下

$$5.267 = \frac{200}{1 + e^{b+3.9}} \quad b = \ln\left(\frac{200}{5.27} - 1\right) - 3.9 = -0.29$$

最终获得参数  $a$ 、 $b$  的初始值分别为 3.9、-0.29。

如果在确定初始值时没有非常明确的范围, 可以先根据对函数的了解设定参数的初始值, 再在约束对话框中设定参数的数值范围。这样可能达到最优回归的步数多一些, 运行时间长一些。只要非参数模型选择正确, 最终总能得到比较满意的结果。

### 2. 调用过程

- (1) 读取数据文件 Data11-05。按 Analyze→Regression→Nonliner 顺序打开主对话框。
- (2) 将变量 *pop* 设置为因变量, 送入 Dependent 框中。
- (3) 在 Model Expression 框中输入估计的模型表达式  $c/(1+2.718^{**}(a+b *deacde))$ 。
- (4) 在 Parameters 框中根据前面计算结果, 设定  $a \approx 3.9$ 、 $b \approx -0.29$ 、 $c=200$ 。
- (5) 在 Save 对话框中选择 Predicted values、Residuals 选项。

### 3. 结果输出见表 11-47~表 11-50, 图 11-42、图 11-43。

表 11-47 是迭代各步的残差平方和与参数  $a$ 、 $b$ 、 $c$  的估计值。每步迭代后, 计算估算值的变化量。表的最后一部分表示在估算完 8 个模型、4 个导数后, 由于两次迭代的最小残差平方和的减少量小于默认的收敛判据 1.E-08 而终止。



表 11-47 每步迭代的残差平方和、参数值

Iteration History <sup>a</sup>				
Iteration Number <sup>a</sup>	Residual Sum of Squares	Parameter		
		a	b	c
1.0	2199.753	3.900	-.290	200.000
1.1	203.656	3.883	-.278	241.492
2.0	203.656	3.883	-.278	241.492
2.1	186.497	3.890	-.279	243.967
3.0	186.497	3.890	-.279	243.967
3.1	186.497	3.889	-.279	243.988
4.0	186.497	3.889	-.279	243.988
4.1	186.497	3.889	-.279	243.987

表 11-48 非线性模型统计量摘要

ANOVA <sup>a</sup>			
Source	Sum of Squares	df	Mean Squares
Regression	123053.531	3	41017.844
Residual	186.497	15	12.433
Uncorrected Total	123240.028	18	
Corrected Total	53293.925	17	

Dependent variable: pop  
a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .997.

根据前面的计算，得出最终的回归方程为

$$y_i = \frac{243.99}{1 + e^{3.89 - 0.28t_i}}$$

表 11-48 为非线性模型统计量的摘要。Sum of Squares 列是统计量的平方和：Regression 是回归平方和，Residual 是残差平方和，Uncorrected Total 为因变量各观测值的平方和，Corrected Total 为因变量各观测值对均值的偏差平方和。

$$R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Corrected Sum of Squares}} = 1 - \frac{186.497}{53293.925} = 1 - 0.0034 = 0.9966$$

表明模型对数据的拟合程度非常好。如果模型的拟合程度非常差， $R^2$  也可能为负值。散点图 11-42 是使用双轴（Dual Axes）图形功能完成。第 2 个纵轴是预测值。从预测值与观测值的散点也可以看出拟合得很好。

注意，不能使用线性模型的检验方法检测非线性模型。即使模型非常正确，残差均值平方也不再是误差方差的无偏估计。为了应用的目的，仍可以比较残差方差和估算总方差，但是  $F$  统计量却不能再用来对假设进行检验。

在非线性模型中不大可能获得每个参数精确的置信区间，大样本一般依靠渐近线的近似值进行估算。表 11-49 为各种参数估计值，表 11-50 为估计参数的渐近相关矩阵。

图 11-43 为残差对观测年代的散点图。观察图形，发现残差的方差随着时间的增长而增长。为计算预测值的渐近标准误和其他统计量，可以进行以残差为因变量的线性回归分析。

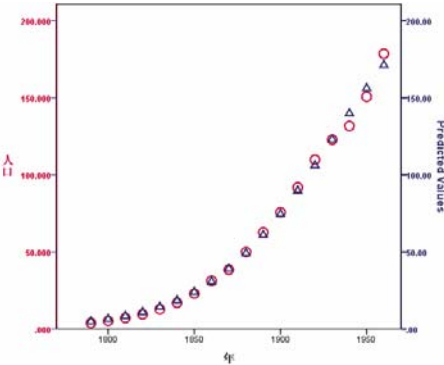


图 11-42 观测与预测值的散点图

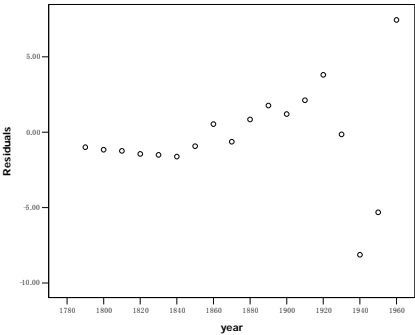


图 11-43 残差一年度散点图

表 11-49 参数估计值

Parameter Estimates				
Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	3.889	.094	3.690	4.089
b	-.279	.016	-.312	-.246
c	243.987	17.968	205.690	282.285

表 11-50 估计参数的渐进相关矩阵

Correlations of Parameter Estimates			
	a	b	c
a	1.000	-.724	-.376
b	-.724	1.000	.904
c	-.376	.904	1.000

如果在表 11-50 中出现非常大的正值或者负值，很可能是由于模型中参数过多（较少参数的模型就能很好地拟合数据），相对来说观测量的数量不足。但不说明模型不适合。

## 11.7 加 权 回 归

### 11.7.1 加权回归的概念

建立线性回归模型的前提是样本方差为一个常数，也就是说所有的观测量在计算过程中具有相同的贡献。这种方法称为正常最小二乘法（OLS）。如果某些观测量的一些特性变异较其他观测量大，此时使用 OLS 方法就不能获得较好的模型。但是，如果它们的变异性是可以通过其他变量进行预测的，就可以使用加权最小二乘法（WLS）来拟合线性回归模型。加权回归给出加权转换的范围，并得出最佳的权数值。

例如，考虑到由于高市值的股票较低市值的股票具有较高的变异性（价格的上下波动），仅使用一般线性回归过程的 OLS 方法进行估算就不能很好地反映通货膨胀与失业率对变异性较大股票价格的影响。而 WLS 方法可以较好地解决这个问题。再诸如健康研究中，各种治疗方法对病人住院时间长短的影响，很明显需要住院时间越长的病情，其表现的变异性就要比住院短的病人的病情所表现的变异性要大。产品研究中，工人的训练水平与产品质量之间的关系，因为产品质量越差，其变异性越大。社会学与犯罪学研究中，犯罪率较高的地区要比犯罪率较低的地区表现出更高的变异性。

#### 1. 诊断与权重估计

##### (1) 图形

参见图 11-44(a)所示的例子（数据 Data11-06 来自 1981 年 DRVPER 和 SMITH），此图中只有两个变量  $x$  和  $y$ 。可以观察到因变量的变异性或者分布随着自变量的增加而增加，这暗示着方差相同的假设已经遭到破坏并且最小平方方法不再是最佳解决方案了。

察看图 11-44(b)预测值与残差散点图，可以得出相同的结论。

##### (2) 估计权重的方法

① 从数据的复制集中估计权重。为了使用加权最小平方方法来估计回归模型，将具有相同特点或近似特点的数据进行编组（数据的复制集）。这样就可以计算因变量相对于每一组具有不同特点的自变量的方差了。此时得到的方差的倒数就是权重。

② 从变量估计权重。如果认为因变量的方差与自变量或者其他变量之间存在关系，

就可以使用 WLS 方法来估计权重。例如，研究收入与受教育程度之间的关系可知那些有研究生学位人员的工资变异要比那些没有获得学位的人员工资的变异高得多。

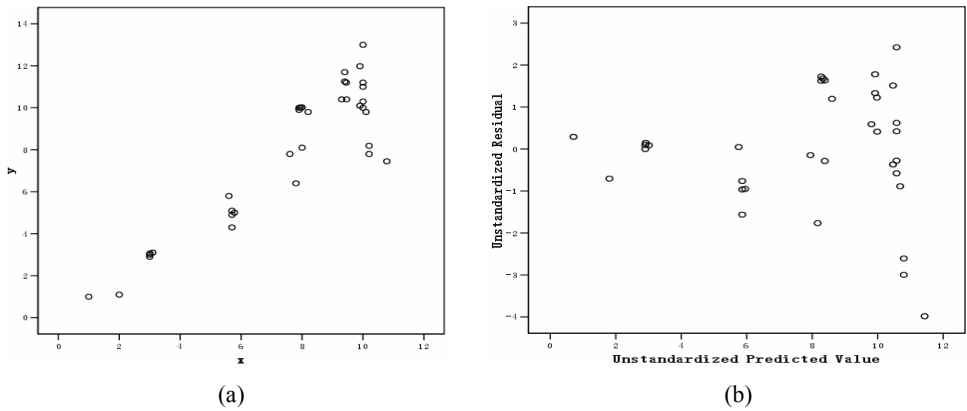


图 11-44 原始变量散点图与残差散点图

方差、变量、指数之间的关系如下： $\text{方差} \propto \text{变量}^{\text{指数}}$ 。可以指定指数值的范围或者一个增量，SPSS 将会在规定的范围内，估计所有指数值的对数似然比值，然后选择出具有最大似然比值的指数值来。

2. 数据要求

自变量和因变量应该是数值型变量，类似宗教、民族和地区这样的分类变量应该被重新编码作为二分（哑）变量或其他类型的对比变量。加权变量必须是与因变量有关的数值型变量。对于自变量的每一个值，要求因变量的分布必须是正态的。因变量和每一个自变量的关系应该是线性的，并且所有的观测量应该是相互独立的。自变量取不同值时，因变量的方差不同，但是这些差异一定是可以根据加权变量预测出来的。

11.7.2 加权回归过程

1. 按 Analyze→Regression→Weight Estimation 顺序打开对话框，见图 11-45。
2. 从左侧的源变量框中选择一个变量作为因变量进入 Dependent 框中。
3. 从源变量框中选择一个或多个的自变量进入 Independent(s)框中作为自变量。
4. 从源变量框中选择一变量，将其选入 Weight Variable 框中，作为加权变量。观测量数据的权重为  $1/wv^{\text{power}}$ ，wv 为加权变量，power 为加权变量指数。
5. 在 Power range 选项中输入将在计算权重的过程中所使用的指数值的范围。在 Power range 后面框中设定初始值，在 through 后面框中设定结束值，在 by 后框中设定步长，应该保证(初始值—结束值)/步长小于等于 150。指数值的范围必须在-6.5 至 7.5 之间。
6. Include constant in equation，模型中包括常数项。
7. 单击 Options 按钮，打开如图 11-46 所示的选择项对话框，确定在数据文件中保存的新变量，并确定方差和估测值的列表形式。

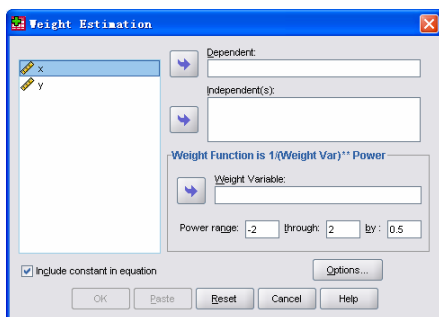


图 11-45 权重估计主对话框

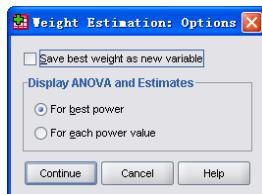


图 11-46 选择项对话框

(1) Save best weight as new variable, 保存新变量的值是根据最大对数似然比函数计算的最佳权重值。变量名为 WGT\_n, n 是运行、生成这个新变量的序号。

(2) Display ANOVA and Estimates 栏, 确定方差和估计值的输出形式。

① For best power, 只输出最终的方差和最佳指数估计值。

② For each power value, 输出方差和所设置范围的指数值。

8. 单击 OK 按钮进行统计分析。

### 11.7.3 加权回归分析实例

【例 9】教学数据 Data11-06 有两个变量  $x$ 、 $y$ 。求以  $x$  为自变量,  $y$  为因本量的回归方程。

1. 打开数据文件后的操作步骤如下:

(1) 按 Analyze→Regression→Weight Estimation 顺序打开 Weight Estimation 对话框。

(2) 选择变量  $y$  为因变量送入 Dependent 框,  $x$  为自变量送入 Independent(s)框。  $x$  变量作为加权变量进入 Weight 框。

(3) 设置加权的指数值, 初始值为 0, 结束值为 2.5, 步长为 0.1。

(4) 在 Options 对话框中选择 Save best weight as new variable, 保存每一个观测量的权重值, 其他选项为默认设置。

(5) 单击 OK 按钮, 提交运算。结果输出见表 11-51~表 11-55。

表 11-51 说明自变量为  $x$ , 因变量为  $y$  及按照 0.1 为步长的权值计算出的对数似然比结果。因为 -55.543526 为最大值, 得到的最佳指数值为 1.900。

表 11-52 为回归效果的统计量, 其中源变量为  $x$ , 因变量为  $y$ , 权重值 1.9。

表 11-53 这些统计量说明模型对数据的拟合程度较好。还要看方差分析结果。

表 11-54 为对回归方程的方差分析。 $F$  值为 567.3, 显著水平值小于 0.01 说明由回归解释的变异远远大于残差可解释的变异, 回归效果是比较好的。

表 11-55 是对回归方程中自变量  $x$  的系数为 0 的假设检验。T 检验的 Sig < 0.05 拒绝  $x$  系数为 0 的假设, 也说明回归效果是好的。所得方程式的最后结果为  $y = -0.283 + 1.077x$ 。

在数据窗中生成新变量WGT\_1。

表 11-51 权值

Log-Likelihood Values <sup>a</sup>		
Power	0	-61.796
	0.1	-61.281
	0.2	-60.775
	0.3	-60.279
	0.4	-59.796
	0.5	-59.327
	0.6	-58.873
	0.7	-58.437
	0.8	-58.020
	0.9	-57.626
	1	-57.255
	1.1	-56.912
	1.2	-56.598
	1.3	-56.319
	1.4	-56.075
	1.5	-55.872
	1.6	-55.714
	1.7	-55.603
	1.8	-55.545
	1.9	-55.544 <sup>a</sup>
	2	-55.603
	2.1	-55.726
	2.2	-55.918
	2.3	-56.182
	2.4	-56.521
	2.5	-56.938

表 11-52 模型描述

Model Description		
Dependent Variable		y
Independent Variables	1	x
Weight	Source	x
	Power Value	1.900

Model: MOD\_1.

表 11-53 模型综述

Model Summary	
Multiple R	.972
R Square	.945
Adjusted R Square	.943
Std. Error of the Estimate	.197
Log-likelihood Function Value	-55.544

表 11-54 方差分析

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	22.116	1	22.116	567.319	.000
Residual	1.286	33	.039		
Total	23.403	34			

表 11-55 模型参数及各种统计量

Coefficients						
	Unstandardized Coefficients		Standardized Coefficients		t	Sig.
	B	Std. Error	Beta	Std. Error		
(Constant)	-.283	.194			-1.457	.155
x	1.077	.045	.972	.041	23.818	.000

2. 与一般线性回归方法进行比较

(1) 以同一数据文件，进行一般线性加权回归分析和一般线性不加权回归分析。

① 一般线性回归：选择变量  $x$  作为自变量， $y$  作为因变量做一次回归。

② 一般线性加权回归：选择  $x$  作自变量， $y$  作因变量，新变量 WGT\_1 作为加权变量送入 WLSWeight 框中。

对比表 11-56 和表 11-57，一般线性回归的的  $R^2$  值为 0.905， $R^2$  为 0.819，明显要小于加权回归的对应值，这说明方程式加权后的效果是十分明显的。

表 11-56 一般线性回归小结

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.905 <sup>a</sup>	.819	.814	1.45535

a. Predictors: (Constant), x

表 11-57 加权线性回归小结

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.972 <sup>a</sup>	.945	.943	.19744

a. Predictors: (Constant), x

(2) 两种回归方法方差分析见表 11-58、表 11-59。系数检验结果见表 11-60、表 11-61。比较表 11-60、表 11-61，发现 WLS 方法和 OLS 方法的斜率(1.077、1.096)和截距(-0.283、-0.387)没有大的差别。但是它们的标准误变化较大：斜率  $b$  的标准误分别为 0.045、0.090，

常数项的标准误分别为 0.194、0.722。

表 11-58 一般线性回归方差分析

ANOVA <sup>b</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	316.433	1	316.433	149.399	.000 <sup>a</sup>
	Residual	69.895	33	2.118		
	Total	386.329	34			

a. Predictors: (Constant), x

b. Dependent Variable: y

表 11-59 加权线性回归的方差分析

ANOVA <sup>a,c</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22.116	1	22.116	567.319	.000 <sup>a</sup>
	Residual	1.286	33	.039		
	Total	23.403	34			

a. Predictors: (Constant), x

b. Dependent Variable: y

c. Weighted Least Squares Regression - Weighted by Weight for y from WLS, MOD\_1  
x\*\* -1.900

表 11-60 一般线性回归的系数检验

Coefficients <sup>a</sup>					
Model		Unstandardized Coefficients		Standardized Coefficients	Sig.
		B	Std. Error	Beta	
1	(Constant)	-.387	.722		.596
	x	1.096	.090	.905	.000

a. Dependent Variable: y

表 11-61 加权回归的系数检验

Coefficients <sup>a,b</sup>					
Model		Unstandardized Coefficients		Standardized Coefficients	Sig.
		B	Std. Error	Beta	
1	(Constant)	-.283	.194		.155
	x	1.077	.045	.972	.000

a. Dependent Variable: y

b. Weighted Least Squares Regression - Weighted by Weight for y from WLS, MOD\_1  
x\*\* -1.900

为了进一步验证加权模型的效果，做转换后的预测值与残差的散点图。需要注意在线性回归过程中首先保存预测值和残差到数据文件中，然后在绘制散点图之前对它们进行转换，转换的方法是它们本身乘以加权变量的 1/2 次方。

绘制的图形，如图 11-47 所示，可以看出转换后预测值对残差值散点图的喇叭形状比图 11-44(b)有了改善。说明 WLS 方法获得了一定的效果。

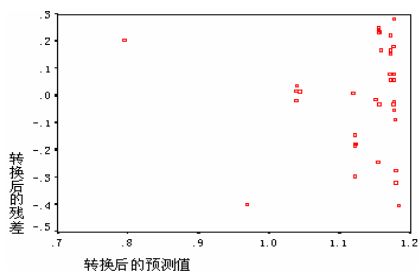


图 11-47 转换后预测值对残差值散点图

## 习 题 11

1. 数据 Data11-10 是某企业 1987~1998 年的经济效益、科研人员、科研经费的统计数据。假定 1999 年该企业科研人员 61 名、科研经费 40 万元，试预测 1999 年该企业的经济效益。

2. 某商场 1989—1998 年的商品流通费用与商品零售额资料如数据 Data11-11 所示。若 1999 年该商场商品零售额为 36.33 亿元，试预测 1999 年该商场商品流通费用。

3. Data11-12 是 R.Norell 进行的一项用电流刺激农场动物的实验数据，其目的是为了求得一成的牲畜对高压电流有反应的临界值。在新农场选址时，要求高压线的辐射电流低于临界值。如果超过，则需要重新选址。试求出临界电流值。

## 第 12 章 非参数检验

习惯上可按下列标准把统计假设划分为两大类：如果假设只对总体分布中的若干个参数指定取值范围，则称这种假设为参数性假设；否则，称为非参数性假设。同样，检验问题也可划分为两大类：在已知总体分布的具体函数形式的前提下，只是其中若干个参数未知，则称这种检验问题为参数检验问题，否则称为非参数检验问题。参数检验问题中的原假设和各择假设都是参数假设，在前几章中我们已经接触到许多这方面的内容。而非参数检验问题的前提中并没有指定总体分布的具体函数形式，故原假设（或称零假设）和各择假设都可以是参数性假设，也可以是非参数性假设。非参数检验是指在总体分布情况不明时，用来检验数据资料是否来自同一个总体的假设的检验方法。由于这些方法一般不涉及总体参数故得名。这类方法的假定前提比参数性假设检验方法少得多，也容易满足，适用于计量信息较弱的资料且计算方法也简便易行，所以得到广泛的应用。

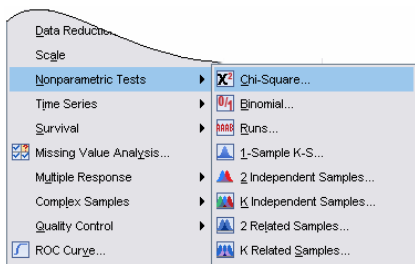


图 12-1 各种非参数检验过程菜单

SPSS 中进行非参数检验由主菜单的 Analyze 下拉菜单中的 Nonparametric Tests 菜单项导出。参见图 12-1。其中包括：Chi-Square Test（卡方检验）、Binomial Test（二项式检验）、Runs Test（游程检验）、1-Sample K-S（一个样本柯尔莫哥洛夫-斯米诺夫检验）、2 Independent Samples Tests（两个独立样本检验）、K Independent Samples（多个独立样本检验）、2 Related Samples（两个相关样本检验）、K Related Samples（多个相关样本检验）。

在上述八种非参数检验方法中，前四种方法通常用来做分布的拟合优度检验，即检验样本所在的总体是否服从某个已知的理论分布。后四种方法通常用于分布位置检验，即检验样本所在的总体的分布位置或形状是否相同。

### 12.1 卡方检验

#### 12.1.1 卡方检验的基本概念

在前面的章节介绍的方法中，往往都事先假定总体服从正态分布，然后对其均值或方差作差异的显著性检验。但某个随机变量是否服从某种特定的分布是需要进行检验的。可以根据以往的经验或实际的观测数据的分布，推测总体可能服从某种分布函数  $F(x)$ ，

利用这些样本数据来具体检验该总体分布函数是否真的就是  $F(x)$ 。卡方检验 (Chi-Square Test) 就是这样一种用来检验给定的概率值下数据来自同一总体的无效假设的方法。通常地, 卡方检验可以用来对分类变量是二项或多项分布的总体作分布的一致性检验。

### 12.1.2 卡方检验过程

1. 按 Analyze→Nonparametric Tests→Chi-Square 顺序展开 Chi-Square Test 对话框, 如图 12-2 所示。

2. 从源变量列表选择一个或多个需要进行检验的变量, 移到 Test Variable List 框中。

3. 在 Expected Range 栏内确定检验值的范围。在默认情况下, 变量的各个截然不同的值被当作分类值。

(1) Get from data, 采用数据中的最小值和最大值所确定的范围。系统默认选项。

(2) Use specified range, 建立指定范围内的分类, 只检验数据中一个子集的值, 在 Lower 和 Upper 参数框中输入检验范围的下限和上限。输入的值须为整数。数据值超过这个指定范围的样品, 不参与分析。

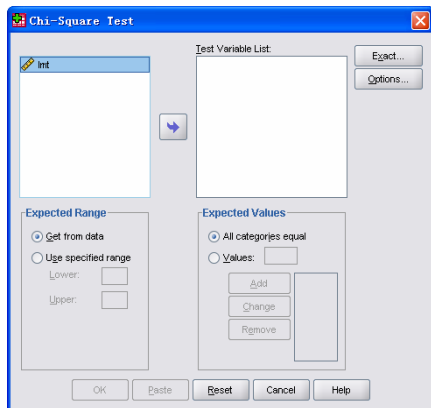


图 12-2 卡方检验主对话框

4. 在 Expected Values 栏中指定期望值

(1) All categories equal, 系统默认的检验是所有组对应的期望值都相同, 这意味着要检验的是总体是否服从均匀分布。

(2) Values, 选定所要检验的是总体是否服从某个给定的分布, 并在其右边的框中输入相应各组所对应的由给定分布计算得的期望值的百分比。该值必须大于 0。并应同原分类次序相同的升序顺序保持一致。这一点非常重要。

每输一个值后按 Add 按钮, 于是在它右边的框中的底部便增加刚输入的期望值百分比, 一直到输完所有的期望值为止。如果输入了错误数据, 选中错误数据, 单击 Remove 按钮即可将它删除。或者输入正确数据, 按 Change 按钮, 则错误值被替换。

5. 单击 Exact 按钮打开如图 12-3 的 Exact Tests 精确检验对话框 (需购买 Exact 选件)。它提供了另外两种计算显著性水平的方法: 精确法和蒙特卡洛 (Monte Carlo) 法。当数据不满足标准渐近法所必须的基本假定条件时, 它们提供了一个获得精确结果的方法。在图 12-3 的对话框中, 共有三个选项可供选择:

① Asymptotic only, 渐近方法, 系统默认选项, 当数据集相当大, 或表格被稠密地填入和完全平衡时, 选择本项。

② Monte Carlo, 本方法与 Exact 都适用于数据集很小, 如样本含量小于 30, 表格



稀疏或不平衡，或样本含量小于 50 且出现小于 5 的期望频数时。选择此项，在 Confidence level 框后输入置信水平，并在 Number of samples 后指定用于此近似法中的样品数量。要想复制结果，每次使用此法时要设置随机数种子。此方法比 Exact 能更快得到结果。

③ Exact，适用条件同 Monte Carlo。选择本项，需要在 Time limit per test 后框中输入最大限制时间。如果检验计算超过设置时限 30 分钟，建议使用 Monte Carlo 法。如果发现没有足够的内存空间，应首先关闭正在运行的任何其他的应用软件，以节省内存供计算使用。如果还不能获得精确结果，应改用 Monte Carlo 法。

6. 单击 Options 按钮，打开 Chi-Square Test: Options 对话框，如图 12-4 所示。

(1) 在 Statistics 栏中选择输出统计量。

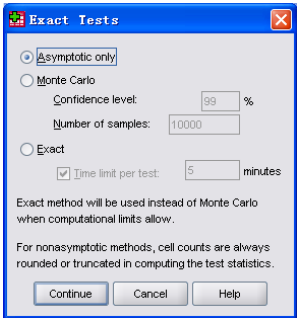


图 12-3 精确检验对话框

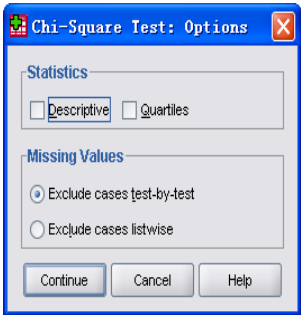


图 12-4 选择对话框

① Descriptive，输出变量的均值、标准差、最大值、最小值和非缺失个体的数量。

② Quartiles，输出四分位数。

(2) Missing Values 栏选择对缺失值的处理方式。

① Exclude cases test-by-test，将参与对比中的缺失值排除在分析之外。

② Exclude cases listwise，剔除任何变量中所有含有缺失值的样品。

7. 单击 OK 按钮，系统立即执行命令。

12.1.3 卡方检验分析实例

【例 1】掷一颗六面体 300 次，结果见表 12-1，取变量名为“lmt”，用数字型数据 1、2、3、4、5、6 分别代表六面的六个点，试问这颗六面体是否均匀。

表 12-1 300 次掷一颗六面体实验观测结果

点数 I	1	2	3	4	5	6
频数 O	43	49	56	45	66	41

(1) 数据录入有两种方式。data12-01 是直接录入原始数据，只有一个变量。在以下应用中可直接使用。如果数据已经整理成表 12-1 形式，应该建立两个变量并按表中方式

录入, 见 data12-01a。这种方式的数据在分析前, 应先用 Data 菜单中的 Weight cases by 过程, 将频数变量定义为权重变量。对变量“六面体[lmt]”进行加权处理。操作参见第 2 章 2.4.1 节中的介绍。在本章的后续部分一律简称为“加权处理”。经加权处理后的变量, 与 data12-01 方式录入的同名变量, 在统计分析中是等价的因此不再加以说明。

## (2) 操作方法

- ① 读取数据文件 data12-01a。
- ② 按 Analyze→Nonparametric Tests→Chi-Square 顺序展开 Chi-Square Test 对话框。
- ③ 选择“六面体[lmt]”变量送入 Test Variable List 框。
- ④ 由于这是一个均匀分布检验, 故直接使用系统默认值, 单击 OK 按钮, 执行运算。

## (3) 输出结果。见表 12-2。

表 12-2 六面体均匀性卡方检验结果

六面体				Test Statistics	
	Observed N	Expected N	Residual		六面体
1	43	50.0	-7.0	Chi-Square <sup>a</sup>	8.960
2	49	50.0	-1.0	df	5
3	56	50.0	6.0	Asymp. Sig.	.111
4	45	50.0	-5.0	a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 50.0.	
5	66	50.0	16.0		
6	41	50.0	-9.0		
Total	300				

说明: 在本例中, 原假设为这枚六面体是均匀分布的。左表的第二列为实际观察值出现的次数, 第三列为在原假设成立的条件下, 理论上点数应出现的期望频数, 第四列为观察频数与理论频数的差值, 即残差。在右表中显示的由上到下依次是统计检验的卡方值、自由度和原假设成立的显著性概率值, 脚注说明的是: 小于 5 的期望频数的单元格为 0, 最小的期望单元格频数为 50。因计算结果  $P=0.111>0.05$ , 故在这个检验中不足以拒绝原假设, 而可以认为这一颗六面体是均匀的。

【例 2】表 12-3 为 100 名健康成年女子血清总蛋白含量, 试问它是否服从均值为 7.36 和标准差为 0.395 的正态分布。如果血清蛋白含量这个变量服从正态分布, 则可用现有样本的均数及标准差作为其所隶属总体的均数及标准差的无偏估计, 通过计算可得, 相应表 12-3 各组的理论期望值的百分比分别为: 6.37%、9.54%、15.67%、20.07%、19.44%、14.64%、8.62%和 5.65%。

表 12-3 100 名健康成年女子血清总蛋白含量表

组 限	6.60	6.80	7.00	7.20	7.40	7.60	7.80	8.00
组内频数	8	8	11	25	24	10	7	7

(1) 两种方式录入的数据分别存放在 data12-02a 和 data12-02 的数据文件中。data12-02a 中的 hb 仿例 1 做法, 作了加权处理。可以打开任意一个数据文件进行分析。

## (2) 操作方法

- ① 按 Analyze→Nonparametric Tests→Chi-Square 顺序展开 Chi-Square Test 对话框。

- ② 选择 hb 变量进入 Test Variable 对话框。
- ③ 在 Expected Values 栏中选择 Values 项，并在 Values 参数框中分别输入 6.37%、9.54%、15.67%、20.07%、19.44%、14.64%、8.62%和 5.65%。在这里值得一提的是，如果输入的百分比总和超过 100%时，系统将重新做归一化处理，因此，在 Values 参数框中输入值时，百分比符号可省略。换句话说，如果已经计算了期望值而非期望值的百分比，则可直接按次序输入期望值，这对统计计算结果和分析没有影响。
- ④ 单击 Options 按钮，在对话框中的 Statistics 栏内选中 Descriptive 和 Quartiles。
- ⑤ 单击 OK 按钮，进行运算。输出结果，见表 12-4。

表 12-4 血清蛋白含量正态分布检验结果

Descriptive Statistics								
血清总蛋白	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
血清总蛋白	100	7.284	.3687	6.6	8.0	7.000	7.200	7.400

Test Statistics			
血清总蛋白			
Chi-Square <sup>a</sup>	6.436		
df	7		
Asymp. Sig.	.490		

<sup>a</sup>. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 5.7.

血清总蛋白			
	Observed N	Expected N	Residual
6.6	8	6.4	1.6
6.8	8	9.5	-1.5
7.0	11	15.7	-4.7
7.2	25	20.1	4.9
7.4	24	19.4	4.6
7.6	10	14.6	-4.6
7.8	7	8.6	-1.6
8.0	7	5.7	1.4
Total	100		

本例中原假设为：血清蛋白含量数据资料服从均值为7.36和标准差为0.395的正态分布。输出表之一给出描述统计量，表头标题依次是样本含量、平均数、标准差、最小值、最大值、第25、50（中位数）和75百分位数。其余两张表的表头说明，参见本章例1的结果。因 $P=0.49>0.05$ ，故在此检验中无法拒绝原假设，可认为血清蛋白含量服从正态分布。

12.2 二项分布检验

12.2.1 二项分布检验的概念与操作

- 1. 二项分布检验的基本概念
- Binomial Test 二项分布检验是用来检验在给定的落入二项式中第一项概率值的前提下数据来自二项分布的无效假设的方法。
- 2. 基本操作
- (1) 按 Analyze→Nonparametric Tests→Binomial 顺序展开如图 12-5 所示的对话框。
- (2) 从变量列表选择一个或多个需要进行检验的变量，移到 Test Variable List 框中。
- (3) 在 Define Dichotomy 栏中定义二分值。
- ① Get from data, 指定的分析变量只有两个有效值，为系统默认选项，无缺失值。
- ② Cut point, 指定的变量超过两个值，要在参数框中输入分界点值，比分界点的值

小的形成第一项，其余的构成第二项。

(4) 在 Test 参数框中指定检验概率值。

系统默认的检验概率是 0.5，这意味着要检验的二项是服从均匀分布的。如果落入每一项中的个体的期望比率不等，换言之，所要检验的二项不是同概率分布，参数框中输入相应第一项所对应的概率期望值。

(5) 选择精确检验方法、输出结果形式及缺失值处理方式，见 12.1.2 节。

(6) 单击 OK 按钮，执行命令。

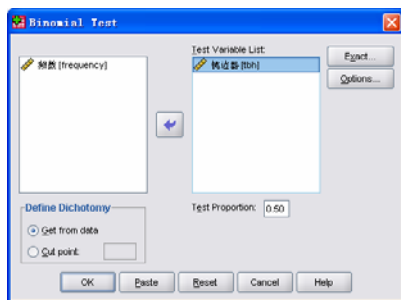


图 12-5 二项分布检验主对话框

## 12.2.2 二项分布检验分析实例

【例 3】掷一枚球类比赛用的挑边器 31 次，出现 A 面或 B 面向上的结果见表 12-5，取变量名为“tbh”，用数字型数据 1 代表“A”，用数字型数据 2 代表“B”，依次在数据库中输入数据。试问这枚挑边器是否均匀。

1. 检验的零假设是，挑边器服从第一项的概率值为 0.5 的二项分布。

表 12-5 31 次掷一枚球类比赛用挑边器实验观测结果

次	1	2	3	4	5	6	7	8	9	10	11	12	13	13	15	16
面	A	B	A	B	B	A	A	A	B	B	A	B	B	A	A	A
次	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
面	B	A	B	B	A	B	B	A	B	A	B	B	A	B	A	

2. 数据文件 data12-03a 或 data12-03。对 data12-03a 中的 tbh 变量进行了加权处理。

3 按 Analyze→Nonparametric Tests→Binomial 顺序展开 Binomial Test 对话框。

4. 选择 tbh 变量进入 Test Variable 对话框。

5. 这是一个均匀分布检验，故直接使用系统默认值。单击 OK 按钮，提交运算。

6. 输出结果，见表 12-6。

表 12-6 挑边器均匀性二项分布检验结果

Binomial Test						
	Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)	
tbh	Group 1	1	.15	.50	1.000 <sup>a</sup>	
	Group 2	2	.16			
	Total	31	1.00			

a. Based on Z Approximation.

本例的原假设为：表中第二列是分类值，第三列为观察频数，第四列为各类的观察概率，第五列为检验概率，最后一列是原假设成立时双侧检验的概率值。因  $P=1.00>0.05$ ，故可认为这一枚挑边器是均匀的。

## 12.3 游程检验

### 12.3.1 游程检验的基本概念

一个游程就是某序列中位于一种符号之前或之后的另一种符号持续的最大主序列,或者说,一个游程是指某序列中同类元素的一个持续的最大主集。

例如,我们做一个掷硬币试验,以概率  $P$  得正面,以概率  $1-P$  得反面,用数字 0 记正面,用数字 1 记反面。不太可能出现多个 0 或多个 1 连续地连在一起,也不太可能 0 和 1 交替频繁地出现。假如做这样的试验 30 次,得到如下的试验纪录

000011100000110000011111100000

如果称连在一起的 0 或连在一起的 1 为一个游程,则上面的例子中有 4 个 0 游程和 3 个 1 游程,共 7 个游程 ( $R=7$ )。记 0 出现的次数为  $n$ ,记 1 出现的次数为  $m$ ,则总的试验次数  $N=n+m$ 。显然,出现 0 或 1 的次数的多少同概率  $P$  有关,但在已知  $n$  和  $m$  时,游程数  $R$  的条件分布就同  $P$  无关了。

游程检验 (Runs Test) 就是根据游程数所作的两分变量的随机性检验。

其原假设为:两分变量有随机性。在原假设成立的前提下,当样本容量很大,即当

$m/n \rightarrow \gamma$  时

$$Z = \frac{R - \frac{2m}{1+\gamma}}{\sqrt{\frac{4\gamma m}{(1+\gamma)^3}}} \rightarrow N(0,1)$$

在给定显著性水平  $\alpha$  后,可用下面的近似公式得到临界值

$$c_1 = \frac{2mn}{m+n} \left[ 1 + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{m+n}} \right] \quad c_2 = \frac{2mn}{m+n} \left[ 1 - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{m+n}} \right]$$

游程检验可用来检验样本的随机性,这对于统计推断是很重要的。游程检验也可用来检验任何序列的随机性,而不管这个序列是怎样产生的。此外还可用来判断两个总体的分布是否相同,从而检验出它们的位置中心有无显著差异。

在具体的实际问题中,并不是所有的数据对都是以 0 或 1 的二元形式来表现的。例如在遇到连续型的计量资料时,可先找出中位数,然后所有的原始数据与中位数来比较,大于中位数的计为 1,小于中位数的计为 0,这样可把计量资料变成一组 0,1 系列。就可按二元变量的随机性方法来做检验了。

如果样本来自的两总体的分布形态存在较大差距,则计算出的游程数会相对比较小。如果游程数比较大,则应是由于两样本数据充分混合的结果,则它们的分布应该不存在显著差异。

### 12.3.2 游程检验过程

1. 按 Analyze→Nonparametric Tests→Runs 顺序单击菜单项, 展开如图 12-6 所示的游程检验主对话框。

2. 从源变量表中选择一个或多个需要进行检验的变量, 移到 Test Variable List 框中。

3. 在 Cut Point 栏内确定划分两类的分割点。在该框中提供了用来定义两类的分割点方法。变量值小于分割点的个体形成第一类, 其他个体形成第二类。可选的分割点如下: Median, 指定中位数做分割点; Mode, 指定众数做分割点; Mean, 指定平均数做分割点; Custom, 将自定义分割点数值输入到后面的编辑框中。

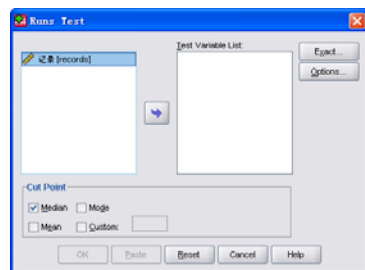


图 12-6 游程检验主对话框

4. 选择精确检验方法、输出结果形式及缺失值处理方式的操作参见 12.1.2 节。

5. 单击 OK 按钮, 执行命令。

### 12.3.3 游程检验分析实例

【例 4】掷硬币 20 次得到的实验数据见表 12-7, 其中 1 表示数字面朝上, 0 表示画面朝上。建立变量 records, 标签为“记录”, 将表 12-7 中的数据录入到 SPSS 中, 存在数据文件 data12-04 中。试问掷硬币试验是否是随机的?

1. 操作步骤

(1) 假设掷硬币的结果是随机的。

(2) 读取数据文件 data12-04。

(3) 按 Analyze→Nonparametric Tests→Runs 顺序展开 Runs Test 对话框。

(4) 选择 records 变量送入 Test Variable List 框中。

(5) 本例中由于 1 出现 11 次, 0 出现 9 次, 中位数和众数都为 1, 故不能选用 Mode (众数) 和 Median (中位数) 作为划分两类的分割点。本例选用 Mean (平均数) 和 Custom (自定义期望值)。分割点值应大于 0 且小于 1, 本例输入 0.5。

(6) 单击 OK 按钮, 提交运算。

2. 输出结果见表 12-8。

3. 输出结果解释

表 12-8 是 Runs Test 游程检验表。它是以 Mean 平均数作为分界点的结果, 计算结果平均数是 0.55; Runs Test2 表是 Custom (自定义值) 0.5 作为分界点的运行结果。

Runs Test 表的第二行起依次为: 检验值为 0.55、小于检验值的样品数为 9、大于等于检验值的样品数为 11、总样品数 20、游程数量 12、Z 值和原假设成立的双侧检验概

率值。因  $P=0.781>0.05$ ，故不拒绝原假设，即掷硬币试验是随机的。

表 12-8 掷硬币试验随机性检验结果

表 12-7 掷硬币 20 次的结果

1	1	0	1	0
0	0	1	1	0
1	0	1	1	1
0	0	1	1	0

Runs Test		Runs Test 2	
	记录		记录
Test Value <sup>a</sup>	.55	Test Value <sup>a</sup>	.50
Cases < Test Value	9	Total Cases	20
Cases ≥ Test Value	11	Number of Runs	12
Total Cases	20	Z	.279
Number of Runs	12	Asymp. Sig. (2-tailed)	.781
Z	.279		
Asymp. Sig. (2-tailed)	.781		
a. Mean		a. User-specified.	

## 12.4 一个样的本柯尔莫哥洛夫-斯米诺夫检验

### 12.4.1 一个样本的柯尔莫哥洛夫-斯米诺夫检验的基本概念

一个样本的柯尔莫哥洛夫-斯米诺夫检验（One-Sample Kolmogorov-Smirnov Test）简称单样本的 K-S 检验。它是用来检验样本来自 Normal 正态、Uniform 均匀或 Poisson 泊松分布总体的假设。这也是一种拟合优度检验方法，它主要是运用某随机变量  $x$  的顺序样本来构造样本分布函数，使得能以一定的概率保证  $x$  的分布函数  $F(x)$  落在某个范围内。

K-S 双侧检验的原假设  $H_0$  为：对所有的  $x$  值  $F(x)=F(x_0)$  成立，备择假设为：至少有一个  $x$  值使  $F(x) \neq F(x_0)$  成立。

设  $S(x)$  表示一组数据的经验分布。定义一组随机样本  $x_1, x_2, \dots, x_n$  的经验分布函数为阶梯函数

$$S(x) = \frac{x_i \leq x \text{ 的个数}}{n}$$

它是小于  $x$  的值的比例，是总体分布  $F(x)$  的一个估计。检验统计量为

$$D = \sup_x |S(x) - F_0(x)|$$

$D$  的分布对一切连续分布  $F(x_0)$  在原假设下是一样的，所以它与分布无关。在实际运算中，由于  $S(x)$  是阶梯函数，只取离散值，所以考虑到跳跃问题，如果有  $n$  个观察值，则可用下面的统计量来代替上面的  $D$

$$D = \max_{1 \leq i \leq n} \left\{ \max(|S(x_i) - F_0(x_i)|, |S(x_{i-1}) - F_0(x_i)|) \right\}$$

当  $n \rightarrow \infty$  时，大样本的渐近公式为

$$P(\sqrt{n}D_n < x) \rightarrow K(x)$$

其分布函数的表达式为

$$K(x) = \begin{cases} 0 & x < 0 \\ \sum_{j=-\infty}^{\infty} (-1)^j \exp(-2j^2 x^2) & x > 0 \end{cases}$$

### 12.4.2 柯尔莫哥洛夫-斯米诺夫检验过程

1. 按 Analyze→Nonparametric Tests→1-Sample K-S 顺序展开如图 12-7 所示对话框。
2. 从左侧变量表中选择一个或多个需进行检验的变量，移到 Test Variable List 框中。
3. 确定要检验的分布。在 Test Distribution 框中，提供了所要检验的分布，分别有 Normal（正态分布）、Uniform（均匀分布）、Poisson（泊松分布）、Exponential（指数分布），系统默认检验正态分布。

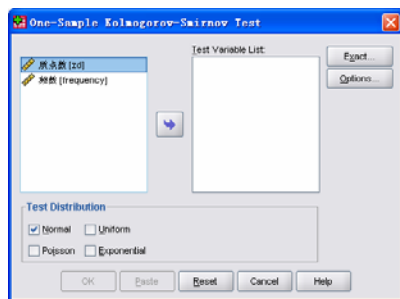


图 12-7 单样本 K-S 检验主对话框

SPSS 命令语言也允许读者指定正态分布的平均数和标准差，为泊松分布指定平均数，为均匀分布指定最大值和最小值。

4. 精确检验的选择、输出结果形式及缺失值处理方式的操作参见 12.1.2 节。
5. 单击 OK 按钮，执行命令。

### 12.4.3 柯尔莫哥洛夫-斯米诺夫检验分析实例

【例 5】数据 data12-05 是卢瑟福与盖革做的一个著名的实验的记录。他们观察由某块放射物质放出的，在 7.5 秒的时间间隔里到达某计数器的  $\alpha$  粒子数。共观察了 2608 次。表 12-9 中变量 zd 记录的是 7.5 秒到达计数器的粒子数，变量 fi 是每个 zd 值出现的次数。试问这种分布规律是否服从泊松分布。

表 12-9 质点实验数据

zd	0	1	2	3	4	5	6	7	8	9	10
frequency	57	203	383	525	532	408	273	139	45	27	16

1. 数据文件 data12-05a 按表中数据录入，定义 frequency 为加权变量。
2. 假设 7.5 秒内到达计数器的粒子数服从泊松分布。
3. 操作步骤
  - (1) 按 Analyze→Nonparametric Tests→1-Sample K-S 顺序单击菜单项，展开对话框。
  - (2) 选择 zd 变量进入 Test Variable 对话框。
  - (3) 在 Test Distribution 框中选中 Poisson，对是否服从泊松分布进行检验。
  - (4) 单击 OK 按钮，提交运算。



4. 输出结果见表 12-10，给出了实验数据的泊松分布参数的计算结果。

表 12-10 中， $N=2608$  为样本量，第 2 行为泊松分布参数的均值，第 3 行是最大极端差异的绝对值，第 4 行为最大极端差异的正数值，第 5 行是最大极端差异的负数值，第 6 行是 K-S Z 值，最后一行是原假设成立的双侧检验的显著性概率。因  $P=0.850>0.05$ ，故不拒绝原假设，无足够证据推翻数据服从泊松分布的假设。

表 12-10 Poisson 检验结果

One-Sample Kolmogorov-Smirnov Test			渐近数
N			2608
Poisson Parameter <sup>a,b</sup>	Mean		3.87
Most Extreme Differences	Absolute		.012
	Positive		.010
	Negative		-.012
Kolmogorov-Smirnov Z			.611
Asymp. Sig. (2-tailed)			.850

a. Test distribution is Poisson.  
b. Calculated from data.

## 12.5 两个独立样本检验

### 12.5.1 两个独立样本检验的用途与基本操作

两个独立样本均服从正态分布时比较均值使用 T 检验。但有时样本所隶属总体的分布类型可能不明或是非正态的，但还是想知道在这种情况下两个独立样本间是否具有相同的分布，两个独立样本检验（Two Independent Samples Test）就是用来处理此类问题的一种有效方法。执行本过程要求的数据文件结构与进行独立样本 T 检验的数据结构一样。

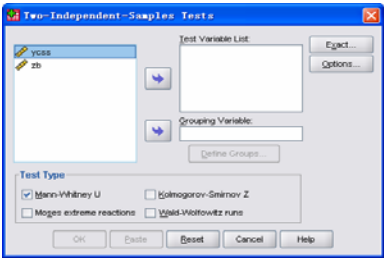


图 12-8 两个独立样本检验主对话框

1. 按 Analyze → Nonparametric Tests → 2 Independent Samples 顺序单击菜单项，展开相应的对话框，如图 12-8 所示。

2. 指定检验变量。从变量列表中选择要进行检验的一个或多个变量，移到 Test Variable List 框中。

3. 指定分组变量名。从左面变量列表中指定用来分组的变量，并移到 Grouping Variable 框中，单击 Define Groups 按钮，展开如图 12-9 所示的对话框，在两个编辑栏中输入分组值。

4. 确定用来进行检验的方法。在 Test Type 框中，提供了可供用来检验的 4 种方法，检验两个独立样本（组）是否来自同一个总体。它们分别是：

(1) Mann-Whitney U，是非常流行的两个独立样本检验，等同于两组的 Wilcoxon 秩和检验和 Kruskal-Wallis 检验。此方法检验两个样本的总体在位置上相等的。来自两组的观察被合并和赋予秩，有结的样品被赋予平均秩。结的数量相对于观察的总数量应该很少。如果总体在位置上是同样的，则秩应该是随机地混合在两个样本里的。换句话说，如果两个样本含量相同，秩和也相同；如果



图 12-9 定义分组对话框

两个样本含量不相同, 则两个样本的平均秩相同。计算第一组得分领先第二组得分的次数和第二组得分领先第一组得分的次数。Mann-Whitney U 统计这两个数量的较少者。还显示较小样本秩和的 Wilcoxon 秩和  $W$  统计量。如果两个样本有相同的观察的数量, 则  $W$  是在 Define Groups 对话框中首先指定的组即 Group1 组的秩和。

(2) Kolmogorov-Smirnov Z, 是更普通的探测两者位置和分布形状上差异的检验。该检验是建立在两个样本的累积分布函数之间的最大绝对差异的基础上的。当这个差异显著地大时, 两个分布被认为是具有差异的。

(3) Moses Extreme Reactions, 假设实验变量影响某个方向上的一些被试对象和相反方向上的其他被试对象。它检验同控制组比较的极端反应。本检验关键是控制组的跨度, 以及当合并控制组时, 实验组里有多少极端值影响跨度。分组对话框里组 1 值定义的是控制组, 来自两个组的观察被合并和赋秩。控制组的跨度是组中的最大和最小值的秩之差加 1。由于跨度范围很容易受偶然因素的影响, 所以两端 5% 的样品被自动地删除。

(4) Wald-Wolfowitz Runs 游程检验是更普通的探测在两者位置和分布形状上差异的检验。该游程检验对两组观察合并和赋予秩。如果两个样本是来自相同的总体, 则两组应被始终随机地分散赋秩。

在这 4 种方法中, 至少应选择一种。系统默认选项为 Mann-Whitney U 法。

5. 选择精确检验、输出结果形式及缺失值处理方式的操作参见 12.1.2 节和图 12-3。

6. 单击 OK 按钮, 提交运算。

## 12.5.2 两个独立样本检验分析实例

【例 6】设有甲、乙两种安眠药, 考虑比较它们的治疗效果, 独立观察 20 名患者。10 人服甲药, 另 10 人服乙药, 睡眠延长的时数见表 12-11。试问这两种药物的疗效有无显著性的差异。

表 12-11 两种安眠药效果对比数据

服甲药者睡眠延长时数	1.9	0.8	1.1	0.1	0.1	4.4	5.5	1.6	4.6	3.4
服乙药者睡眠延长时数	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0

延长的睡眠时数的分布情况不明, 因此用非参数检验的方法。

数据文件 data12-06 中, 变量 ycss 为服药后睡眠时间延长的时数; zb 变量值为实验组别。组别用 1 表示服乙药、2 表示服甲药。录入数据后按变量 ycss 值降序排序。

1. 假设两组药物对延长睡眠时间的无显著差异。实际就是检验这两个独立样本是否具有相同的分布。

2. 操作步骤

(1) 按 Analyze→Nonparametric Tests→2 Independent Samples 顺序展开相应对话框。

(2) 选择 ycss 变量进入 Test Variable List 框中。

(3) 选择 zb 变量进入 Grouping Variable 框中。单击 Define Groups 按钮, 展开 Define

Groups 对话框，在 Group 1 框中输入 1，在 Group 2 框中输入 2。

(4) 由于在录入数据后已对数据作了排序处理，故在 Test Type 框中可选择其中的任何一种方法。本例选择了 4 种方法。

(5) 单击 OK 按钮，提交运算。输出结果见表 12-12~表 12-15。

表 12-12 Mann-Whitney U 检验结果

Ranks				
组别	N	Mean Rank	Sum of Ranks	
睡眠延长时数	10	7.75	77.50	
服甲药组	10	13.25	132.50	
服乙药组	10			
Total	20			

Test Statistics <sup>a</sup>	
Mann-Whitney U	22.500
Wilcoxon W	77.500
Z	-2.095
Asymp. Sig. (2-tailed)	.036
Exact Sig. (2*(1-tailed Sig.))	.035 <sup>b</sup>

a. Not corrected for ties.  
b. Grouping Variable: 组别

表 12-13 Moses 检验结果

Frequencies	
组别	N
睡眠延长时数	10
服甲药组 (Control)	10
服乙药组 (Experimental)	10
Total	20

Test Statistics <sup>a,b</sup>	
Observed Control Group Span	17
Trimmed Control Group Span	15
Outliers Trimmed from each End	1

a. Moses Test  
b. Grouping Variable: 组别

表12-14 两样本 Kolmogorov-Smirnov Z检验结果

Frequencies		
组别	N	
睡眠延长时数	10	
服甲药组	10	
服乙药组	10	
Total	20	

Test Statistics <sup>a</sup>	
Most Extreme Differences	Absolute .500
	Positive .500
	Negative .000
Kolmogorov-Smirnov Z	1.118
Asymp. Sig. (2-tailed)	.164

a. Grouping Variable: 组别

表12-15 Wald-Wolfowitz检验结果

Frequencies	
组别	N
睡眠延长时数	10
服甲药组	10
服乙药组	10
Total	20

Test Statistics <sup>a,c</sup>			
睡眠延长时数	Number of Runs	Z	Exact Sig. (1-tailed)
Minimum Possible	6 <sup>b</sup>	-2.068	.019
Maximum Possible	10 <sup>b</sup>	-.230	.414

a. There are 2 inter-group ties involving 7 cases.  
b. Wald-Wolfowitz Test  
c. Grouping Variable: 组别

因 4 种方法计算的  $P$  值除 Mann-Whitney U 检验外均大于 0.05，故可认为这两种药物的疗效无显著性差异。

12.6 多个独立样本检验

12.6.1 多个独立样本检验的用途与操作

上面所提到的两个独立样本检验是多个独立样本检验中最基本的形式，要解决多个独立样本间是否具有相同的分布的问题，需借助于多个独立样本检验（Test for Several Independent Samples）方法。操作方法如下：

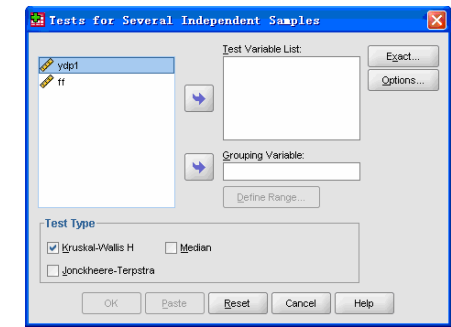


图 12-10 多个独立样本检验主对话框

1. 按 Analyze→Nonparametric Tests→K Independent Samples 顺序展开如图 12-10 所示的对话框。
2. 指定检验变量。  
从左侧变量列表选择一个或多个需要进行检验的变量，移到 Test Variable List 框中。
3. 指定分组变量值范围。  
从变量列表中选择分组变量移到 Grouping Variable 框中，单击 Define Range 按钮，在相应的对话框中，定义分组变量值范围。

4. 确定用来进行检验的方法有 3 种:

(1) Kruskal-Wallis H 检验是 Mann-Whitney U 检验的扩展, 类似单方向方差分析, 探究分布位置上的差异。该方法假设从  $k$  个无序的总体中抽取样本, 是系统默认的方法。

(2) Median (中位数) 检验, 是很普通的检验, 但效率不高, 探究在位置和形状上的分布差异。该方法假设从  $k$  个无序的总体中抽取样本。

(3) Jonckheere-Terpstra 检验。当  $k$  个总体有序 (升序或降序) 时, 此检验方法非常有效。例如,  $k$  个总体可以描述  $k$  个增加的温度。检验的假设是不同的温度产生同样反应的分布, 备择假设: 温度升高反应剧烈。这里, 假设两个样本是有序的, 因此, 使用 Jonckheere-Terpstra 检验是最适当的。安装了 Exact Tests 时此检验才是可选用的。

5. 选择精确检验、输出结果形式及缺失值处理方式的操作参见 12.1.2 节。

## 12.6.2 多个独立样本检验分析实例

【例 7】某车间用四种不同的操作方法各做若干批试验, 试验中优等品率 (%) 数据资料见表 12-16, 试问操作方法对产品的优等品率是否有显著影响。

表 12-16 四种不同操作方法的优等品率试验数据

试验批号	操作方法 1	操作方法 2	操作方法 3	操作方法 4
1	12.1	18.3	12.7	7.3
2	14.8	49.6	25.1	1.9
3	15.3	10.1	47.0	5.8
4	11.4	35.6	16.3	10.1
5	10.8	26.2	30.4	9.4
6		8.9		

1. 数据在文件 data12-07 中, 变量: ydp1 为优等品率、ff 为操作方法。

2. 操作步骤

(1) 按 Analyze→Nonparametric Tests→K Independent Samples 顺序展开主对话框。

(2) 选择 ydp1 变量进入 Test Variable List 框中。

(3) 选择 ff 变量进入 Grouping Variable 框中。单击 Define Range 按钮, 展开 Define Range 对话框, 在 Minimum 框中输入 1, 在 Maximum 框中输入 4。

(4) 在 Test Type 框选择 Kruskal-Wallis H 法。因每组观测值数量太少 Median 法不适用, 故不选择。

(5) 单击 OK 按钮, 提交运算。

3. 输出结果, 见表 12-17。检验结果包括两个小表。

在 Kruskal-Wallis H 法中, 计算的  $P$  值等于 0.009, 小于 0.05, 故可认为这四种不同的操作方法对产品优等品率是有显著影响的。

表 12-17 Kruskal-Wallis 检验结果

Ranks				Test Statistics <sup>a,b</sup>	
操作方法	N	Mean Rank	Chi-Square	df	优等品率 %
1	5	10.40	11.530	3	
2	6	13.75			
3	5	15.80			
4	5	3.50			
Total	21				

a. Kruskal Wallis Test  
b. Grouping Variable: 操作方法

## 12.7 两个相关样本检验

### 12.7.1 两个相关样本检验的用途与操作

在实际的研究工作中，我们经常会遇到从同一个被测试对象身上测试两个或多个观测值的情况，这样的数据间就不再是相互独立的了，而是彼此相关。在此种情况下，检验样本间是否具有相同的分布，要用两个相关样本检验 Two-Related Samples Tests。

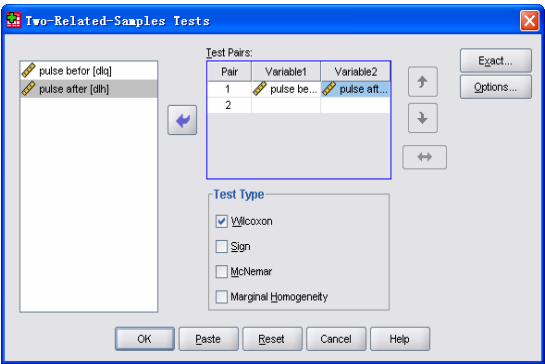


图 12-11 两个相关样本主对话框

在 Test Type 框中，提供了 4 种方法。

(1) 如果数据是连续的，使用 Sign test 符号检验或 Wilcoxon Signed-Rank test 威尔科克森符号等级检验。

① 符号检验对所有样品计算两个变量值间的差并将差值分为正、负或结 3 类。如果两个变量有类似的分布，则正、负数差异无显著不同。

② 威尔科克森符号等级检验要求较小比较的两个变量分布形状相似，不但考虑两个符号的差异，且考虑成对样品数值之间差异。所以比符号检验更有效。

(2) 如果数据是二元的，则使用 McNemar 检验。每个被试对象的反应被抽查两次，分别在指定事件发生的前、后。McNemar 检验确定初始的反应的比率（事件前）是否等于最终反应的比率（事件后）。本检验用来探究取决于实验干涉前、后设计中反应的变化。

(3) 如果数据是分类的，则使用边缘同质检验 Marginal Homogeneity test。它是 McNemar 检验从二元反应向多重反应的扩展。它使用卡方分布检验实验干涉前、后设计中反应的变化。如果安装了 ExactTest 附件时，边缘同质检验才有效。

- 4. 精确检验方法的选择、输出结果形式及缺失值处理方式的选择操作见 12.1.2 节。
- 5. 单击 OK 按钮，提交运算。

1. 按 Analyze→Nonparametric Tests →2 Related Samples 顺序展开如图 12-11 所示的对话框。

2. 指定检验变量对  
从左面变量列表中同时选择两个待检验的变量，送到 Test Pairs 框中，在 Pair 的 Variable 1 和 Variable 2 中依次出现所选择的两个变量名。如果相关的成对变量为多组，则重复上述操作。

3. 确定检验方法  
本节里检验比较两个相关变量的分布。检验的方法根据数据类型确定。

## 12.7.2 两个相关样本检验分析实例

【例 8】为研究长跑运动对增强普通高校学生的心功能效果，对某校 15 名男生进行测试，经过 5 个月的长跑锻炼后看其晨脉是否减少。锻炼前后的晨脉数据见表 12-18。试问锻炼前后的晨脉间有无显著性的差异。

表 12-18 长跑锻炼后晨脉变化

锻炼前	70	76	56	63	63	56	58	60	65	65	75	66	56	59	70
锻炼后	48	54	60	64	48	55	54	45	51	48	56	48	64	50	54

1. 数据文件 data12-08 中，变量 dlq 为锻炼前的晨脉，变量 dlh 为锻炼后的晨脉。

2. 操作步骤

(1) 按 Analyze→Nonparametric Tests→2 Related Samples 顺序展开主对话框。

(2) 选择 dlq 和 dlh 变量进入 Test Pairs 框中。

(3) 在 Test Type 框中选中 Wilcoxon 和 Sign。

(4) 单击 OK 按钮，提交运算。

3. 输出结果，见表 12-19 和表 12-20。

因两种检验方法计算的  $P$  值均小于 0.05，故可认为锻炼前后晨脉间有显著性的差异。

表 12-19 Wilcoxon Signed Ranks 检验结果

Ranks				
	N	Mean Rank	Sum of Ranks	
锻炼后晨脉 - 锻炼前晨脉	12 <sup>a</sup>	9.17	1.10E2	
Positive Ranks	3 <sup>b</sup>	3.33	10.00	
Ties <sup>c</sup>	0 <sup>c</sup>			
Total	15			

a. 锻炼后晨脉 < 锻炼前晨脉

b. 锻炼后晨脉 > 锻炼前晨脉

c. 锻炼后晨脉 = 锻炼前晨脉

Test Statistics <sup>a</sup>	
Z	-2.842 <sup>a</sup>
Asymp. Sig. (2-tailed)	.004

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

表 12-20 Sign 检验结果

Frequencies	
锻炼后晨脉 - 锻炼前晨脉	N
Negative ...	12
Positive Differences <sup>a</sup>	3
Ties <sup>b</sup>	0
Total	15

a. 锻炼后晨脉 < 锻炼前晨脉

b. 锻炼后晨脉 > 锻炼前晨脉

c. 锻炼后晨脉 = 锻炼前晨脉

Test Statistics <sup>a</sup>	
Exact Sig. (2-tailed)	.035 <sup>a</sup>

a. Binomial distribution used.

b. Sign Test

## 12.8 多个相关样本检验

### 12.8.1 多个相关样本检验的用途与操作

两个相关样本检验是多个相关样本检验最基本的形式，要解决多个相关样本间是否具有相同的分布的问题，使用多个相关样本检验 Test for Several Related Samples 方法。

1. 按 Analyze→Nonparametric Tests→K Related Samples 顺序展开如图 12-12 所示的对话框。

2. 从源变量列表中选择需要进行检验的一个或多个变量，移到 Test Variables 框中。

3. 在 Test Type 栏中，根据变量类型确定检验方法。

(1) Friedman 是等同于一个样本重复测定设计或每单元一个观察值的双向方差分析的非参数检验。此检验的无效假设： $k$  个相关的变量来自同一个总体。对每个样品的  $k$

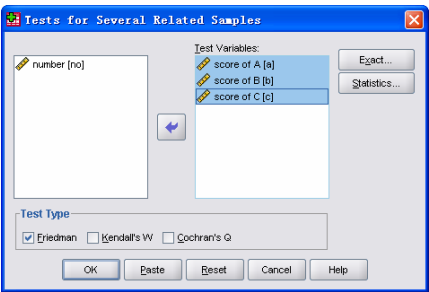


图 12-12 多个独立样本检验主对话框

$k$  个样本情形的 McNemar 检验的扩展。Cochran's Q 检验假设几个相关的两分变量有相同的均数。变量是在同一个个体或在配对个体上测定的。

- 4. 选择精确检验方法、输出结果形式及缺失值处理方式的操作参见 12.1.2 节。
- 5. 单击 OK 按钮提交系统执行。

12.8.2 多个相关样本检验分析实例

【例 9】某商店想了解顾客对几种款式不同的衬衣的喜爱程度。某日询问了 9 名顾客，请他们对 3 种款式的衬衣按喜爱程度排次序（最喜爱的给秩 1，其次的给秩 2，再次的给秩 3），结果见表 12-21，试问顾客对 3 种款式的衬衣的喜爱程度是否相同。数据在 data12-09 中。变量  $a$ ,  $b$ ,  $c$  的数据分别是各顾客对款式 A、B、C 衬衣的喜爱程度。

表 12-21 顾客对不同款式衬衣喜爱程度

顾客号	1	2	3	4	5	6	7	8	9
款式 A	2	2	2	1	3	1	2	1	1
款式 B	3	3	3	3	2	2	3	3	3
款式 C	1	1	1	2	1	3	1	2	2

- 1. 假设顾客对 3 种款式衬衣喜爱程度无显著差异。
- 2. 按 Analyze→Nonparametric→K Related Samples 顺序展开相应的对话框。选择  $a$ ,  $b$ ,  $c$  变量进入 Test Variables 框中。在 Test Type 框中选择 Friedman 方法和 Kendall's W 方法。单击 OK 按钮，提交运算。
- 3. 输出结果，见表 12-22 和表 12-23。

表 12-22 Friedman 检验结果

Ranks		Test Statistics <sup>a</sup>	
	Mean Rank	N	
A款得分	1.67	Chi-Square	8.222
B款得分	2.78	df	2
C款得分	1.56	Asymp. Sig.	.016

a. Friedman Test

表 12-23 Kendall's W 检验结果

Ranks		Test Statistics	
	Mean Rank	N	
A款得分	1.67	Kendall's ...	.457
B款得分	2.78	Chi-Square	8.222
C款得分	1.56	df	2
		Asymp. Sig.	.016

a. Kendall's Coefficient of Concordance

因两种检验方法计算的  $P$  值均等于 0.016 小于 0.05，故可认为顾客对 3 种款式的衬

个变量值被从 1 到  $k$  赋值。根据这些秩进行统计检验。

(2) Kendall's W 是标准化的 Friedman 统计量，是调和系数。它是比率之间一致性的测度。每个样品是一个鉴定人或定价人，每个变量是一个条件或被鉴定的人。对每个变量，计算秩和。Kendall's W 范围在 0（不同意）和 1（同意）之间。

(3) Cochran's Q 同 Friedman 检验是相同的，适用于所有的应答是二元的情况。它是对

衣的喜爱程度是不相同的。

## 12.9 非参数假设检验过程的命令语句

### 1. 非参数假设检验的命令语句清单

```

NPAR TESTS [CHISQUARE=varlist[(lo,hi)]] [/EXPECTED={EQUAL} {f1,f2,...,fn}]
           [/K-S({UNIFORM [min,max]} {NORMAL [mean,stddev]} {POISSON [mean]}
            {EXPONENTIAL [mean]} )=varlist]
           [/RUNS({MEAN} {MEDIAN} {MODE} {value})=varlist]
           [/BINOMIAL[(.5) {p}]=varlist[{value1,value2} {value}]]
           [/MCNEMAR=varlist [WITH varlist [(PAIRED)]]]
           [/SIGN=varlist [WITH varlist [(PAIRED)]]] [/WILCOXON=varlist [WITH varlist [(PAIRED)]]]
           [/MH=varlist [WITH varlist [(PAIRED)]]]†† [/COCHRAN=varlist] [/FRIEDMAN=varlist]
           [/KENDALL=varlist] [/M-W=varlist BY var (value1,value2)] [/K-S=varlist BY var (value1,value2)]
           [/W-W=varlist BY var (value1,value2)] [/MOSES[(n)]=varlist BY var (value1,value2)]
           [/K-W=varlist BY var (value1,value2)] [/J-T=varlist BY var (value1,value2)]††
           [/MEDIAN[(value)]=varlist BY var (value1,value2)]
           [/MISSING=[{ANALYSIS**}] {LISTWISE}] [INCLUDE] [/SAMPLE]
           [/STATISTICS={DESCRIPTIVES} [QUARTILES] [ALL]]
           [/METHOD= {MC[CIN({99.0} {value})]}
           [SAMPLES({10000} {value})] ]†† {EXACT [TIMER({5} {value})]}

```

\*\*表示所在子命令省略时的系统默认值。††表示安装Exact Tests任选项时有效。

### 2. NPAR TESTS命令

NPAR TESTS命令可以进行的有效的检验根据数据的组织形式分成三种类型：一个样本检验分析一个变量；相关样本检验和独立样本检验比较同一样品集的两个或多个变量。独立样本检验分析两个由分组变量值确定的独立样本。

一次调用NPAR TESTS命令，STATISTICS，SAMPLE和MISSING三个子命令只能各指定一次，检验子命令可多次指定。最多使用100个子命令，合计最多500个变量。子命令出现的顺序任意。字符串变量不能出现在该命令中，否则该命令停止执行。

### 3. 子命令

(1) BINOMIAL子命令检验二分变量的观测值分布同指定的二项分布有相同的期望。关键字后面首先给出第一类样品期望比例 $p$ ，省略此项，默认 $p=0.5$ 。等号后连接变量表。在变量名后括号中给出该变量的两个合法值。给出单个值认为是分隔点，样品值等于或小于分隔点形成第一类，其余为第二类。默认每个变量只有两个值，指定的每个变量是与样品期望的比例为 $p$ 的二项分布进行比较，默认输出包括每组的有效样品数，检



验的比例和观察比例的双侧概率。如果检验比例是默认的0.5,显示双侧概率。否则显示单侧概率。单侧检验的方向取决于在第一类里观察的比例。如果观察比例远大于检验比例,则报告第一类观察值显著地更多。如果观察的比例小于或等于检验比例,则报告第一类观察显著地更少。检验总是在观察方向上进行。

(2) CHISQUARE 子命令在关键字后用等号连接变量表,可以在变量后的括号中给出该变量值最大值和最小值表示的范围。范围外的样品不参与分析。还可以在 EXPECTED 子命令中指定期望比例。NPAR TESTS 根据变量分类的观察频数和期望频数间的差异计算卡方统计量。系统默认各个分类里的期望频数相等。输出包括频数分布、期望频数、残差、卡方、自由度和概率。子命令中最多 200 个值。

EXPECTED 子命令按分类值的升序指定不相等的期望频数,默认各分类值期望频数相等,指定 EQUAL 同省略该子命令效果同相。EXPECTED 适用于在 CHISQUARE 子命令中指定的所有变量。可以 CHISQUARE 和 EXPECTED 子命令配合使用多次。

不指定检验变量的范围,则对每个变量产生 Chi-Square Frequency 表。如果指定范围,则建立适于范围内每个值的整数值分类。非整数分类值被截尾。对所有指定变量产生合并 Chi-Square Frequency 表。期望值被解释为比例,而非绝对值。值被总计,每个值的频数都除以总数计算在相应分类中样品期望比例。

(3) COCHRAN 子命令在关键字后面用等号连接两个或多个编码相同的二分变量表。对这些相关的二分变量进行 Cochran's Q (科克伦 Q) 检验,检验它们的值是否有同样分布。输出 Cochran Frequencies 表中显示每个变量的频数分布、样品数、科克伦 Q 统计量、自由度和在 Test Statistics 表里的概率。对  $k$  个二分变量创建  $k \times 2$  (变量 $\times$ 分类) 列联表,计算各个变量的比例。计算比较所有变量的单个检验。科克伦 Q 统计近似服从卡方分布。

(4) FRIEDMAN 子命令在关键字后面用等号连接两个或多个有序的变量。它检验  $k$  个相关样品是否来自相同的总体。每个样品  $k$  个变量的值从 1 到  $k$  被赋秩,计算每个变量上所有样品的平均秩。检验统计量近似服从卡方分布。在输出的 Friedman Ranks 表里,显示各个变量的平均秩、有效样品数、卡方、自由度和在 Test Statistics 表里的概率。对各单个变量的检验结果进行比较。

(5) J-T 子命令,安装了 Exact Tests 时有效。J-T 子命令执行 Jonckheere-Terpstra 检验,它检验由分组变量定义的  $k$  个独立样品是否来自相同的总体。当  $k$  个总体具有自然分类时,本检验特别有效。J-T 子命令在关键字后面用等号连接要检验的两个或多个有序变量的变量表。关键字 BY 后面给出一个分组变量并在圆括号中给出一对值,成对值定义的范围里的每个值形成一个组。有范围外值的样品不参与分析。子命令对在 BY 前指定的各变量计算 Jonckheere-Terpstra 统计量,它近似服从正态分布。如果指定/METHOD 子命令,并且总体数  $k$  大于 5,则  $P$  值是用 Monte Carlo 抽样方法估计的。当  $k$  超过 5 时,精确的  $P$  值无用。单侧推论的方向是由标准检验统计量的标记指出。输出显示分组变量的水平数、样品总数、观察值、标准值、平均数和统计检验的标准差、双侧渐近显著性

意义。如果指定/METHOD 子命令, 则显示单侧和双侧精确或 Monte Carlo 概率。

(6) K-S 子命令 (One-Sample), 在关键字 K-S 后面括号中指定检验的分布关键字和分布参数, 可选分布和关键字是: 正态分布 NORMAL、泊松分布 POISSON、指数分布 EXPONENTIAL 和一致分布 UNIFORM。后面用等号连接要检验的变量表。

K-S 单样本检验对指定变量计算检验统计量。用变量的累积分布函数同指定的分布进行比较。输出显示有效值数量、检验分布的参数、最极端绝对值、正和负差值、各变量的双侧概率和 Kolmogorov-Smirnov Z 值。该值计算观察值和检验分布函数间最大绝对值差。K-S 概率标准假定检验分布是完全预先指定的。当检验分布的参数是从样本估计时, 检验统计量的分布和作为结果的概率是不同的, 不做校正。对 100000 或更大的平均数, 使用近似泊松分布的正态分布。

(7) K-S 子命令(Two-Sample), 关键字后面等号连接变量表, BY 连接一个分组变量和圆括号里的两个值, 每个值对应一组样品, 其顺序决定最大正数和最大负数的差异。

K-S 检验由分组变量定义的两个独立变量的分布否是相同。对在 BY 前的各变量计算检验统计量。比较两组观察值的累积分布、最大正数、负数和绝对差。对两个分布的中位数、离差、偏斜度等方面的任何差异, 检验是灵敏的。输出 Frequency 表, 显示各组中有效样品数和最大绝对值、两组间正和负的差值、Kolmogorov-Smirnov Z 和在 Test Statistics 表里各个变量的双侧概率。

(8) K-W 子命令, 别名 Kruskal-Wallis。关键字后面等号连接待分析的变量表, BY 连接一个分组变量并在圆括号中指定确定范围的两个值。范围中的每个值对应一组样品,  $k$  个值形成  $k$  个独立样本。子命令对在 BY 前各变量计算各样品的秩和各组的秩和。计算检验统计量 Kruskal-Wallis H, 它近似服从卡方分布。检验由分组变量定义的  $k$  个独立样本是否来自于同一个总体。输出显示有效样品数和在 Ranks 表中各组变量的平均秩和卡方值、自由度及在 Test Statistics 表里输出假设检验的概率。

(9) KENDALL 子命令在关键字后面等号连接待分析的变量表。KENDALL 检验  $k$  个相关样本是否来自同一总体。如果变量表包括  $k$  个变量, 子命令对各个样品,  $k$  个变量的值从 1 到  $k$  排秩, 对各变量计算平均秩。对结进行校正。计算 Kendall's W 和相应的卡方统计量。W 测度每个样品对各变量的评价等级的一致性, 其值的范围在 0 (不一致) 和 1 (完全一致) 之间。输出的 Ranks 表包括各变量的平均秩、有效样品数、Kendall's W、卡方、自由度; 在 Test Statistics 表中输出假设检验的概率。

(10) M-W 子命令关键字后面等号连接变量表, 关键字 BY 后面连接一个分组变量和在圆括号中的两个值, 把样品分为两组。计算检验统计量  $U$  (组 1 得分领先组 2 得分的次数), 用各样品的秩去检验各组是否来自同一个总体。如果样品含量小于等于 40, 计算精确的显著性水平。否则  $U$  被转换成正态分布  $Z$  统计量, 并计算正态近似  $P$  值。输出各组的: 有效样品数、变量的平均秩。在 Ranks 表里给出秩和、Mann-Whitney U、Wilcoxon W (较小组的秩和)、 $Z$  统计量。在 Test Statistics 表中给出假设检验的概率。

(11) MCNEMAR 子命令检验配对的两个二分变量间的合并值是否是等可能的。关键字后面用等号连接两个编码相同的二分变量, 可用 WITH 连接另一个变量表。不使用 WITH, 两个变量组成一对; 使用 WITH, 其前后变量数应该相同。在第二个变量表后面的圆括号中可以使用关键字 PAIRED, 前后顺序相同的变量配成一对。不使用 WITH 不能指定 PAIRED。在计算检验统计量时, 只考虑有两个变量是不同的值的合并。如果样品含量小于 25, 改变从第一个变量到第二个变量的值, 则使用二项分布来计算概率。输出包括每对变量的交叉表、所有成对变量的有效样品数、卡方和在 Test Statistics 表中显示各对变量的假设检验的概率。

(12) MEDIAN 子命令检验  $k$  个独立样本是否来自具有相同中位数的总体。在关键字后面等号连接变量表, BY 后连接一个分组变量并在圆括号里给出两个值。如果  $\text{value1} < \text{value2}$ , 则在两个值定义的范围内, 每个值形成一个组, 执行  $k$  个样本的检验。如果  $\text{value1} > \text{value2}$ , 形成两个组, 执行两个样本检验。可在关键字后的括号中指定中位数, 否则该子命令根据检验里的所有样品计算中位数。用大于中位数和小于或等于中位数的样品数量的计数构造  $2 \times k$  列联表。对 BY 前的各变量计算检验统计量。若样品含量  $N \geq 30$ , 计算卡方统计量。 $N < 30$  时, 使用费雪精确双侧检验替代卡方检验。对各个变量输出 Frequency 表, 显示各分类里大于、小于或等于中位数样品数和有效样品数、中位数、卡方、自由度及在 Test Statistics 表中显示各对变量的假设检验的概率。

(13) MH 子命令是从二分变量的 McNemar 检验到多重的扩展。(安装 Exact Tests 时有效) 除要求关键字后面用等号连接的是两个包含多个值的变量外, 其余规定与子命令 MCNEMAR 的语法相同。对两个变量含有不同值的样品进行边际同质性检验。

本检验这是在重复测定情形下经常使用的方法。检验两个多项式表的等同性, 即两个成对的顺序变量间不同值的合并值是否是等可能的。显示各个非对角单元中的计数。每个表格的第一行用总体 1 指定分类选择, 第二行用总体 2 指定分类选择。用第一行得分的总和计算检验统计量。输出显示所有检验变量截然不同值的数量、有效的非对角单元计数的数量、平均数、标准差、检验统计量的观察和标准值、每个成对变量的近似双侧检验概率。如果指定/METHOD 子命令, 则显示单侧和双侧精确或 Monte Carlo 概率。

(14) MOSES 子命令在关键字后面括号中指定尾部要剔除的样品数 (非百分比), 默认按递增排序剔除尾部 5%。用等号连接待检验的顺序测度的变量, BY 后面是分组变量, 圆括号中给出定义实验组和对照组两个值。该子命令进行极值反应检验, 计算实验组的跨度、最小与最大实验值之间的样品数, 对有结的样品不做修正。从而检验顺序变量的范围是否在实验组和对照组上是相同的。输出显示各变量样品数量总计和在各组里的样品数量的频数表; 显示被剔除的奇异值的数量, 以及剔除奇异值前、后实验组的跨度; 显示 Statistics Test 表包括有、无奇异值的跨度的单侧检验的概率。

(15) RUNS 子命令将待检验变量按变量值大于等于或小于分割点值将样品分为两组, 检验二分变量值的顺序是否是随机的。输出包括检验值即分割点的 Run Test 表, 游

程数量, 小于、大于或等于分割点的样品的数量及对各个变量的检验统计  $Z$  的双侧概率。

在关键字后的括号中给出分割点值或从均值 MEAN、中位数 MEDIAN 或众数 MODE 中选择一个作为分割点。在括号后等号连接待检验的变量表。

(16) SIGN 子命令在关键字后面给出顺序测度的待分析变量表, 其余语法规定与 MCNEMAR 相同。SIGN 检验在两个相关样本里的两个成对变量的分布是否是相同的, 计算成对变量值间正和负差值的计数, 忽略结。如果观察到的差异小于等于 25, 则采用二项分布计算概率。否则, 用  $Z$  分布计算概率。在对大样本含量的无效假设下,  $Z$  近似地服从均数为 0、标准差为 1 的正态分布。输出显示 Frequency 表, 包括正差值对的数量、负差值对的数量、结的数量和总数, Test Statistics 表包括  $Z$  统计量和双侧检验的概率。

(17) W-W 子命令, 关键字后面等号连接分析变量表, BY 后面给出分组变量, 在圆括号中给出两个定义分组的值。该子命令先合并两个组包含有结的样品, 并按升序排列, 计算最小和最大可能的游程。对分析变量计算检验统计量。检验其分布是否与两个独立样本相同。对容量小于或等于 30 的样品, 计算精确单侧概率。否则使用近似正态分布的方法。输出 Frequency 表, 包括分析变量的有效样品总数和每组有效样品数、游程数; Test Statistics 表包括  $Z$  和  $Z$  的单侧概率。如果有结, 则显示游程可能的最小和最大数量、它们的  $Z$  统计量和单侧概率。

(18) WILCOXON 子命令语法规定与 MCNEMAR 相同。该子命令检验两个相关样本中的成对变量的分布是否是相同。计算成对变量间的正负差值的计数、绝对差秩、正秩和与负秩和, 并从正、负秩和中计算检验统计量  $Z$ 。在大样本含量无效假设下,  $Z$  近似地服从均数为 0、标准差为 1 的正态分布。输出 Ranks 表, 包括有效样品的对数、正和负差异, 各自的均数和秩和以及结的数量; 显示 Test Statistics 表, 包括  $Z$  和  $Z$  的概率。

(19) STATISTICS 子命令在关键字后面等号连接汇总统计量关键字, 要求为在 NPAR TESTS 命令中指定的变量计算汇总统计量。汇总统计量关键字: DESCRIPTIVES 输出包括平均数、最大值、最小值、标准差和在命令中指定的各变量的有效样品数; QUARTILES 四分位数, 输出包括第 25、50 和 75 百分位数和样品的数量; ALL 包括所有统计量。

(20) MISSING 子命令指定含有缺失值样品的处理方法。LISTWISE 分析时剔除任何子命令中指定的变量含有缺失值的样品; ANALYSIS 在执行各子命令时, 只剔除参与分析的变量中含有缺失值的样品, 是系统默认的; INCLUDE 读者缺失值作为有效值处理。

(21) SAMPLE 子命令随机抽取内存里存储样品组成样本。抽样对游程检验无效。

(22) METHOD 子命令 (安装了 Exact Tests 时有效) 为需要的统计显示附加的结果。默认显示标准渐近的结果。如果指定了权重变量, 所有方法用最接近的整数权重计算。选项如下:

① MC 对所有的统计根据 Monte Carlo 抽样方法显示无偏估计和置信区间。还显示渐近结果。要计算精确结果 MC 失效。CIN( $n$ )对 Monte Carlo 估计指定置信水平。仅当指定/METHOD=MC 时, CIN 才有效。CIN 的默认值为 99.0。可以指定一个 0.01~99.9

之间的数, 包括 0.01 和 99.9, 作为置信区间。

② SAMPLES 在计算 Monte Carlo 精确估计的  $P$  值时, 指定样本数量。大样本容量可导致狭窄的置信限, 且计算时间更长。指定范围 1~1000000000 之间的任何整数作为样本容量。默认值为 10000。

③ EXACT 除渐近结果外, 为所有统计计算精确显著性水平。如果指定 EXACT 和 MC 两个关键字, 则只提供精确结果。计算精确的  $P$  值可以使内存高度集中。如果指定 /METHOD=EXACT, 并发现计算结果时内存不足, 则首先应关闭正在运行占用很多有效内存的其他应用软件。还能扩大交换文件的容量。如果仍不能获得精确结果, 指定 /METHOD=MC 去获得精确  $P$  值的 Monte Carlo 估计。如果选择 /METHOD=EXACT, 则任选项 TIMER 关键字是有效的。

④ TIMER( $n$ ) 为各统计指定允许运行精确分析的最大分钟数。如果达到时限, 则检验被终止, 不提供精确结果。仅当指定 /METHOD=EXACT 时, TIMER 才可用。可以为 TIMER 指定任何整数值。为 TIMER 指定 0 值, 完全地关闭计时器。TIMER 的系统默认值为 5 分钟。如果一个检验超过 30 分钟时限, 则推荐使用 Monte Carlo, 胜于精确法。

## 习 题 12

1. 什么是非参数检验? SPSS 的哪个过程可进行非参数检验? 共包括几种方法?
2. 100 名健康成年女子血清蛋白含量 data12-02, 试用 Chi-Square 过程检验健康成年女子血清蛋白含量是否服从正态分布?
3. 对一台设备进行寿命试验, 记录 10 次无故障工作时间, 并从小到大排列在数据文件 data12-10 中。问此设备的无故障工作时间是否服从指数分布?
4. 一个监听装置收到的信号, 记录在 data12-11 中。能否说该信号是纯粹随机干扰?
5. 两个地点的地表土壤 PH 值记录在 data12-12 中。问这两个地点的平均 PH 值是否一样。
6. 10 个病人在进行了某种药物疗法前后血压 (单位: 毫米汞柱收缩压) 记录在 data12-13 中。问药物疗法是否有效?
7. data12-14 是某村 20 个村民对 4 个候选人 (A、B、C、D) 的赞同与否的调查 (数字 1 表示赞同, 0 表示不赞同), 试用 Cochran's Q 法检验村民是否对这 4 个候选人有不同的看法?

# 第 13 章 聚类分析与判别分析

## 13.1 聚类、判别分析及其分析过程

分类学是人类认识世界的基础科学。聚类分析和判别分析是研究事物分类的基本方法，广泛地应用于自然科学、社会科学、工农业生产各个领域。

### 13.1.1 聚类分析

聚类分析（Cluster Analysis）是根据事物本身的特性研究个体分类的方法。聚类分析的原则是同一类中的个体有较大的相似性，不同类中的个体差异很大。

根据分类对象的不同，分为样本聚类和变量聚类。

#### 1. 样本聚类

样本聚类在统计学中又称为 Q 型聚类。用 SPSS 的术语来说就是对事件 Cases（或称样品或称观测量）进行聚类。是根据被观测的对象的各種特征，即反映被观测对象的特征的各种变量值进行分类。例如用 K-Means 聚类分析，可以根据观众对电视机外观偏好的特点把电视机外观分为  $k$  组，并把该结果用于确定营销市场的分类，或把被调查的城市进行分类，以便对不同城市的策略进行比较。

应该注意的是，不同的目的选用不同的指标作为分类的依据。例如为选拔少年运动员所选用的指标，就不同于课外活动小组所选用的指标。对啤酒按价格进行分类和按成分进行分类所选用的指标也是不同的。

#### 2. 变量聚类

变量聚类在统计学中又称为 R 型聚类。反映同一事物特点的变量很多，我们是根据所研究的问题选择部分变量对事物的某一方面进行研究。由于人类对客观事物的认识是有限的，往往难以找出彼此独立的有代表性的变量，而影响对问题的进一步认识和研究。例如在回归分析中，由于自变量的共线性导致偏回归系数不能真正反映自变量对因变量的影响。因此往往先要进行变量聚类，找出彼此独立且有代表性的自变量，而又不丢失大部分信息。在生产活动中也不乏需要进行变量聚类的实例。制衣业制定衣服型号就是根据人体各部分尺寸数据找出最有代表性的指标如身長、胸围、裤长、腰围作为上衣和裤子的代表性指标。制鞋业中制定鞋的型号也是如此。变量聚类使批量生产成为可能。

无论哪种聚类分析得出的结论都是为了某种目的所做的工作，往往并非在自然界真实存在这样的类。

13.1.2 判别分析

判别分析是根据表明事物特点的变量值和它们所属的类，求出判别函数，根据判别函数对未知所属类别的事物进行分类的一种分析方法。

在自然科学和社会科学的各个领域经常遇到需要对某个个体属于哪一类进行判断。例如，动物学家对动物分类的研究往往需要获得某个动物属于哪一科、目、纲等的判断，就可以根据判别分析已经得出的判别函数进行判断。

判别分析必须已知样品的所属类别。如果类别未知，必须先做聚类。根据样本聚类的结果进行判别分析，得出判别函数，进而对其他研究对象属于哪一类做出判断。例如在选拔少年运动员时，首先要根据已经有的少年运动员的身体形态、身体素质、心理素质、生理功能的各种指标（变量）进行测试，得到各种指标的测试值（变量值），据此对少年运动员进行分类。根据分类结果再求出选材的判别函数，作为选材的依据。又如，可以根据啤酒中含有的酒精成分、钠成分及所含热量“卡路里”数值对啤酒进行分类。

判别分析与聚类分析不同点在于，判别分析要求已知一系列反映事物特征的数值变量的值，并且已知各个体的分类。

13.1.3 聚类与判别分析过程

SPSS 中进行聚类分析和判别分析的统计分析过程，是由 Analyze 菜单中的 Classify

命令导出的。如图 13-1 所示。二级菜单中是进行聚类分析、判别分析的过程清单，其中包括：

- 1. TwoStep Cluster 两步聚类是一个探索性的分析工具，可以分析大数据文件并自动确定最好的分析结果。
- 2. K-Means Cluster 快速聚类分析过程，仅对观测量进行快速聚类。
- 3. Hierarchical Cluster 分层聚类，进行样本聚类
- 和变量聚类的过程。
- 4. Tree 聚类树过程。
- 5. Discriminate 进行判别分析的过程。

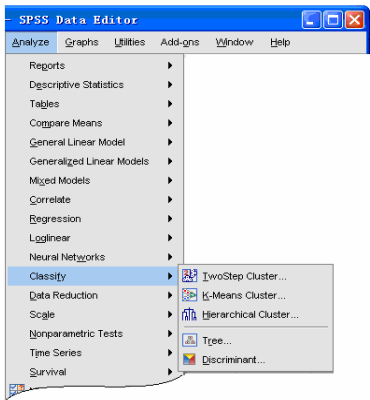


图 13-1 各种聚类分析过程

13.2 两 步 聚 类

13.2.1 两步聚类概述

1. 两步聚类的概念

TwoStep Cluster 过程是一个探索性的分析工具，为揭示自然的分类或分组而设计。

这个过程所使用的算法有几个特色区别于传统的聚类分析技术。其特点是：分类变量和连续变量都可以参与两步聚类分析；该过程可以自动确定分类数；可以高效率地分析大数据集；读者可以自己设置用于运算的内存容量。

两步聚类法在聚类过程中除了使用传统的欧式距离外，为了处理分类变量和连续变量，TwoStep Cluster 过程用似然距离测度，它要求模型中的变量是独立的。分类变量是多项式分布，连续变量是正态分布。虽然经验表明，参与分析的变量违反这一假设的情况下有时也可以得出结果，还是应该使用其他 SPSS 过程检验参与分析的变量是否符合分类变量和连续变量在分布方面的要求。

可以使用两个变量的相关过程 (Bivariate Correlations) 去检验两个连续变量之间的独立性。使用交叉表 (Crosstabs) 过程检验两个分类变量之间的独立性。使用 Means 过程检验连续变量和分类变量之间的独立性。用 Explore 过程检验连续变量的正态性。使用 Chi-Square Test 过程检验分类变量是否是多项式分布的。

所谓两步聚类就是，第一步对每个观测量考察一遍，确定类中心。根据相近者为同一类的原则，计算距离并把与类中心距离最小的观测量分到相应的各类中去。这个过程称作构建一个分类的特征树 (CF)。首先，它把一个观测量放在树的叶节点根部，该节点含有该观测量的变量信息。然后，使用距离测度作为相似性的判据，每个后续的观测量根据它与已经存在的节点的相似性归到某类中去。如果相似则将该观测量加在一个已经存在的节点上，形成该节点的树叶，而如果不相似，就形成一个新节点。

第二步，使用凝聚算法对特征树的叶节点分组。凝聚算法可用来产生一个结果范围。为确定最好的类数，对每一个聚类结果使用 BIC 判据或 AIC 判据作为聚类判据进行比较，得出最后的聚类结果。

两步聚类过程的输出提供聚类得出结果的类数判据 (AIC、BIC)、聚类最终结果的类频数等各类变量的描述统计量，可以产生类频数条形图、类频数饼图和变量重要性图。

## 2. 有关的术语

(1) Cluster Features (CF) Tree, 聚类特征树。在聚类的第一步，根据计算的距离确定类结构。每类有一个节点，属于该类的观测量就是该节点的树叶。由于树叶的不断加构成树枝。第一步聚类过程就是 CF 树成长的过程。

(2) AIC 或 BIC 是在聚类的第二步凝聚过程中用到的两个判据，是两个算法即 Akaike (AIC) 判据或贝叶斯判据 (BIC)。

(3) Tuning the Algorithm, 调谐算法。两步聚类过程可以自动进行聚类，也可以人为控制聚类过程。在人为控制情况下，自己指定参数，在这里称作调谐 (Tuning)。参数指定了，CF 树的规模就基本确定了。

(4) Noise Handling, 噪声处理。由于两步聚类要处理大数据集，在构建 CF 树时，如果指定了类数和算法的参数，例如一个 CF 树最多的分枝数，一个叶节点最大子节点数等，那么在第一步聚类过程中，当观测量很多时，就可能 CF 树满了，不能再长了。没



有在树上的观测量就称为噪声。对这些待处理的观测量，读者可以调整算法参数，让 CF 树能容纳更多的观测量，将其保留在某类中或者丢掉。这种处理称作噪声处理。

(5) Outlier 局外者。根据噪声处理参数聚类结束时被丢掉的观测量称作局外者，单独构成一类，不计在聚类结果的类数中。

13.2.2 两步聚类过程

在使用两步聚类分析过程之前，应该对各变量的变量测度类型在 Variable Viewer 窗口中进行认真的定义。并对各变量的独立性和分布特征进行检验。

两步聚类的操作步骤如下：

1. 建立或读入数据文件后，按 Analyze→Classify→TwoStep Cluster 顺序单击鼠标，展开如图 13-2 所示的对话框。

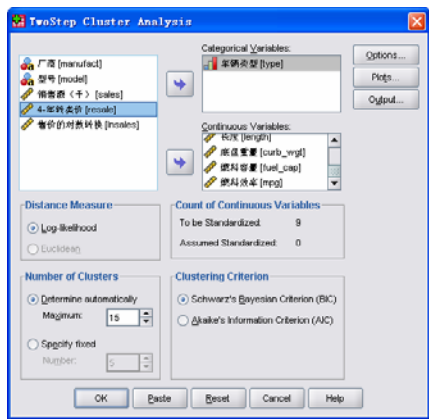


图 13-2 两步聚类分析主对话框

2. 在主对话框中：

(1) 指定分析变量。在左面的源变量栏中显示了可以参加两步聚类分析的两种类型的变量。

选择参与聚类分析的分类变量，单击上面的向右箭头按钮，将其移到右面的 Categorical 矩形框中；选择连续型变量，单击下面的向右箭头按钮，将其送入 Continuous Variables 框中。

(2) Distance Measure 栏，选择计算两类间的相似程度的算法。

① Log-likelihood，该算法要求所有变量彼此独立，连续变量是正态分布的，分类变量是多项式分布的。

② Euclidean，欧式距离法测度两类之间的“直线”距离。当所有参与聚类的变量都是连续变量时此方法才适用。

(3) 确定类数。在 Number of Clusters 栏中指定要聚成几类。其中有两个单选项。

① Determine automatically 自动确定类数，两步聚类过程用在 Clustering Criterion 判据组中指定的判据，自动确定最好的类数。在 Maximum 框中输入一个正整数，指定该过程应该考虑的最大类数。默认的最大类数是 15。最后的聚类结果，类数在 1 至指定的最大类数之间。

② Specify fixed，在 Number 框中输入一个正整数作为要求聚成的固定的类数。最后聚类结果必须是指定的类数。

(4) Count of Continuous Variables 栏，显示连续变量的计数。即在 Options 对话框中指定要进行标准化 to be standadized 的连续变量个数和假设已经标准化的连续变量的个数 Assumed standadized。

(5) Clustering Criterion 栏, 指定确定类数的判据。

① Schwartz's Bayesian Criterion, 可以指定 Schwartz 施瓦茨的贝叶斯判据 (BIC)。

② Akaike's Information Criterion, 可以指定 Akaike 信息判据 (AIC)。

3. Options 在主对话框中单击 Options 按钮打开

如图 13-3 所示对话框。

(1) Outlier Treatment 栏, 选择在特征树满时, 对局外者观测量继续加入特征树的处理方法。如果类特征树满了, 该组选项允许在聚类时对待分类的观测量作特殊处理使 CF 树是完整的, 如果不能接受更多的观测量, 在叶节点和非叶节点处可以分开。

Use noise handling, 在 Percentage 后面给出一个百分比。如果某节点包含的观测量数与最大叶子数之比小于指定的百分比, 就被认为是叶子稀少。当把观测量放到叶子稀少处, CF 树会长大。在树再

次长大后, 如果可能, 待分类的观测量将会被放进 CF 树。否则, 会被当作局外者丢弃。

如果不选择这个选项, 聚类结束后, 那些不能被指派到任何一类中的观测量形成局外类。

(2) Memory Allocation 栏, 允许指定一个聚类过程中使用的最大存储空间 (MB)。如果两步聚类运行时需要占用的空间超出了该最大值, 会使用磁盘存储内存中放不下的信息。默认的容量是 64MB。可以指定一个大于等于 4MB 的数值, 或者请教系统管理员后再确定这个数值。这是处理和分析大样本所需要的。如果这个值太低, 计算法则可能找不到正确的或者希望的类数。

(3) Standardization of Continuous Variables 栏, 聚类算法要求连续变量先完成标准化。任何连续变量都作为要被标准化的变量列在右边 To be Standardized 矩形框中。选择已经事先标准化的变量, 单击向左箭头按钮, 将其送入左面的 Assumed Standardized 矩形框中, 而要被标准化的变量留在右边框中。可以对连续变量事先进行标准化, 以便节省聚类过程所花费的计算时间, 并简化操作。

(4) 高级选项。单击 Advanced Options 按钮打开高级选项对话框如图 13-4 所示。

① CF Tree Tuning Criteria 栏, 聚类算法设置聚类特征 (CF) 树的特殊性, 应该谨慎地改变。

- Initial Distance Change Threshold 框, 初始距离变化极限。这是用于增长 CF 树的起始极限。如果要把一个给定的观测量插入到 CF 树的一个节点上, 产生的紧密性值应该比初始值要小。该叶子就不会被断开。如果密度值超过了初始值, 叶子会被断开生成分支形成节点。系统默认值为 0, 即开始时两个观测量一定是各为一类。

- Maximum Branches (per leaf node) 框, 每个节点的最大分枝数。也就是一个节点可

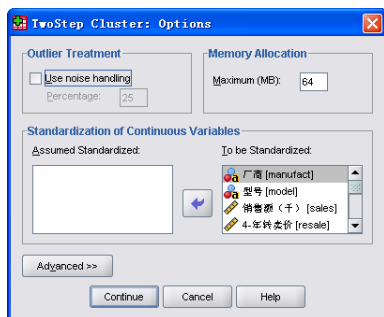


图 13-3 两步聚类选项对话框

以具有的最大子节点数。系统默认值为 8。Maximum Tree Depth (levels)框, 最大树深度, 即 CF 树节点可以有的最大水平数。

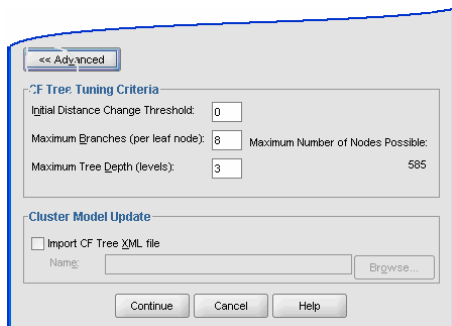


图 13-4 高级选项对话框

• **Maximum Number of Nodes Possible**, 最大可能的节点数 585。这是指可以由该分析过程产生的潜在的 CF 树节点的最大数, 由公式  $(b^{d+1} - 1) / (b - 1)$  计算。这里的  $b$  是最大分枝数,  $d$  是最大树深度。根据系统默认值可以计算出默认的节点数为 585。最低限度, 每个节点需要 16 个字节。一个很大的 CF 树能极大占用和消耗系统资源并反过来影响分析过程的执行。因此在这个对话框中要小心设置各参数。

② **Cluster Model Update** 栏, 聚类模型更新。选择 **Import CF tree XML file**, 允许引入并用当前的数据文件修改以前生成的原聚类模型。引入的文件是 XML 格式的 CF 树。在主对话框中指定分析变量的顺序必须与以前分析时指定的变量顺序相同。除非明确地把新模型信息写到相同的文件名下, 否则 XML 文件保持不变。

如果要更新模型, 使用产生原模型时指定的与 CF 树有关的选项。更明确地说, 使用生成原模型所用的距离测度、噪声处理、存储器设置或 CF 调谐判据等的设置, 而不用在当前对话框中所选项和设置的参数。

注意: 当对一个原聚类模型进行修改时, 该过程假设在当前的工作数据文件中没有被选择的观测量用于产生原分类模型。另外还假设, 用于模型修改的观测量与用于产生原模型的观测量来自相同的总体。也就是说, 两个数据集中的同名的连续变量的均值和标准差假设是相等的; 分类变量的水平都相同。如果新的和旧的观测量集来自不同的总体, 为了得到最好的结果, 应该根据两个数据集的组合来运行两步聚类分析。

4. 统计图选项。在主对话框中单击 **Plot** 按钮, 打开绘图对话框如图 13-5 所示。

(1) **Within cluster percentage chart**, 对每个分类变量, 提供聚类条形图, 表明各类中分类变量各水平的频数。同时显示表明每个连续变量在各类内的变异的图。

(2) **Cluster pie chart**, 显示各类构成比的饼图, 显示各类中观测量数和百分比。同时显示表明每个连续变量在各类内的变异的图。

(3) **Variable Importance Plot** 栏选择每个变量在各类中的重要性图。复选项有:

① **Rank of variable importance**, 确定是否对每个类或每个变量产生图。选择了该项, 激活 **Rank Variables**:

• **By cluster**, 是默认选项。对每变量输出一个重要性图, 每个图比较一个变量在各类中的重要性, 按类排序。

• **By variable**, 对每类输出一个重要性图, 要求在每类中比较各变量的重要性, 按重

要性排序。

② Importance Measure 栏, 选择测度变量重要性的统计量。

- Chi-square or t-test of significance, 用 Pearson 卡方作为分类变量重要性的测度, 用  $t$  作为连续变量重要性的测度。

- Significance, 对连续变量进行均值相等的检验, 给出  $1-p$  值, 对分类变量给出期望频数。

③ Confidence level, 设置置信水平, 是针对变量在类内的分布与变量的整体分布相等的假设检验设置的置信水平。指定一个小于 100 大于等于 50 的数。如果按变量 (by variable) 生成重要性图或者画出显著性, 置信水平对应的统计量的临界值在变量重要性图中显示成一条垂直线。在产生条形图或饼图同时产生的反映各连续变量在每类中变异的高低图中, 此置信水平决定每个高低线的高度。

④ Omit insignificant variables, 忽略无关紧要的变量。那些在指定的置信水平上不显著的变量不显示在变量重要性图中。

5. 输出选项。在主对话框中单击 Output 按钮, 打开对话框, 如图 13-6 所示。

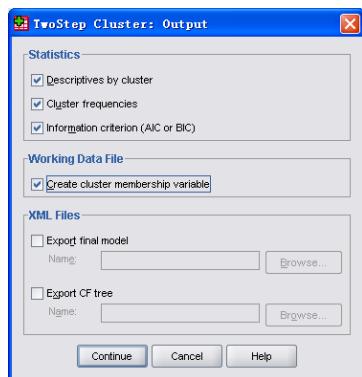


图 13-6 两步聚类 Output 对话框

(2) Working Data File 栏, 保存变量到工作数据文件。

Create cluster membership variable, 产生类成员变量。该变量包括每个观测的类标识号, 变量名是  $tsc\_n$ , 这里  $n$  是正整数, 表明该工作数据文件保存的变量在给定的期间的完成顺序。

(3) XML Files 栏, 以 XML 格式输出最终的聚类模型和 CF 树。

① Export final model, 最终聚类模型要输出到指定的文件。

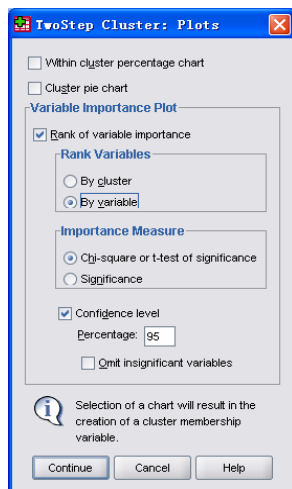


图 13-5 Plot 对话框

(1) Statistics 栏, 该组提供显示聚类结果的选项。对最终的聚类模型, 产生描述统计量和类频数。

① Descriptives by cluster, 显示两个表描述每一类中的变量。在一个表中, 对连续变量分类给出均值、标准差。另一个表分类给出频数。

② Cluster frequencies, 给出每类观测量数的表。

③ Information criterion (AIC or BIC), 根据在主对话框中所选择的判据, 对不同的类数显示一个包括 AIC 或 BIC 值的表格。该表仅在类数是自动确定时才提供。如果类数是固定的, 该设置被忽略, 不输出该表格。



/NUMCLUSTERS AUTO 15 BIC	④
/HANDLENOISE 0	⑤
/MEMALLOCATE 64	⑥
/CRITERIA INITHRESHOLD (0) MXBRANCH (8) MXLEVEL (3)	⑦
/PLOT VARCHART COMPARE BYVAR CONFIDENCE 95	⑧
/PRINT IC COUNT SUMMARY	⑨
/SAVE VARIABLE=TSC_8871.	⑩
AIM TSC_8871	(2)
/CATEGORICAL type	①
/CONTINUOUS price engine_s horsepower wheelbas width length curb_wgt fuel_cap mpg	②
/PLOT ERRORBAR IMPORTANCE (X=VARIABLE Y=TEST)	③
/CRITERIA ADJUST=BONFERRONI CI=95 SHOWREFLINE=YES HIDENOTSIG=NO .	④

### 程序说明

(1) TWOSTEPCLUSTER 语句调用两步聚类过程。以下是子命令：

① CATEGORICAL VARIABLES 子命令定义参与分析的分类变量 type。

② CONTINUOUS VARIABLES 子命令定义参与分析的连续变量，等号后是变量名。  
此处为 9 个连续变量。

③ DISTANCE 子命令设置距离算法为 LIKELIHOOD。

④ NUMCLUSTERS 子命令设置两步聚类采用 BIC 判据，AUTO 关键字设定要自动确定类数，最大类数设置为 15，即要求最后的聚类结果在 1~15 类之间。

⑤ HANDLENOISE 子命令设置处理观测量加入到特征树的方法，括号中的 0 表明不采用噪声处理法。

⑥ MEMALLOCATE 子命令设置进行聚类过程中使用的内存容量为 64MB。

⑦ CRITERIA 子命令设置进行聚类时的特征树的特性。关键字 INITHRESHOLD 设置初始距离变化极限为 0。MXBRANCH 关键字设置一个节点可以具有的最大子节点数为 8。MXLEVEL 关键字设置最大树深度即 CF 树节点可以有的最大水平数为 3。

⑧ PLOT 子命令设置要求输出的统计图，关键字 VARCHART 要求生成变量重要性图，COMPARE BYVAR 要求按变量重要性值排序，关键字 CONFIDENCE 设定置信水平，数值 95 表示 95%。

⑨ PRINT 子命令设置对输出信息的要求，关键字 COUNT 要求输出各类计数，关键字 SUMMARY 要求输出连续变量的描述统计量，对分类变量输出各水平频数期望值。

⑩ SAVE 子命令要求生成新变量，关键字 VARIABLE 后边是系统给定的变量名。

(2) AIM 语句调用生成新变量并作统计图的过程，其后的 TSC\_8871 是生成在当前数据文件中的新变量名。每运行一次给出一个不同的变量名。其值为各观测量所属类别。

① CATEGORICAL 子命令指定 type 是参与分析的分类变量名。

② CONTINUOUS 子命令指定 price、engine\_s、horsepow wheelbas、width length、curb\_wgt、fuel\_cap、mpg 是参与聚类分析并在统计图中出现的连续变量。

③ PLOT 子命令指定生成的统计图。表示各类中的观测量数所占的百分比；IMPORTANCE 关键字指定生成变量重要性图。

④ CRITERIA 子命令指定聚类判据。ADJUST=BONFERRONI 指定使用贝叶斯判据；CI=95 指定置信水平；SHOWREFLINE=YES 指定显示参考线；HIDENOTSIG=NO 指定重要性不显著的变量不在图中出现。

6. 运行结果见表 13-2~表 13-5 和图 13-7~图 13-9。

(1) 表 13-2 自动聚类表概括了按照所选择的类数聚类的过程。从左至右：

第一列是系统默认的对聚类为 1 类到 15 类运算的次序号。

第二列对每个可能的类数计算聚类判据。BIC 的值越小表明模型越好，在这种情况下，最好的聚类结果 BIC 值最小。但是，BIC 值也会随着类数的增加而不断减少，这时就要用 BIC 的变化值和距离测度的变化值来确定最好的聚类结果。一个好的结果应该有相当大的 BIC 变化率和大的距离测度的变化率。

第三列 BIC Change 是当前的 BIC 值减去前一个 BIC 值的差。即 BIC 的变化值。

第四列 Ratio of BIC Changes 是当前 BIC Change 与前一个的比值，即 BIC 值变化率。

第五列为距离测度的变化率。本例按

BIC 最小准则取最后的聚类结果，为 3 类。聚为 3 类时的 BIC 变化率和距离变化率也都相当大。

(2) 表 13-3 是类分布表。表明了每个类中的频数。观测量总数为 157 个，5 个由于

表 13-2 自动聚类过程

Auto-Clustering				
Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change <sup>a</sup>	Ratio of BIC Changes <sup>b</sup>	Ratio of Distance Measures <sup>c</sup>
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	.367	2.190
4	897.559	11.635	-.048	1.368
5	931.760	34.201	-.142	1.036
6	968.073	36.313	-.151	1.576
7	1026.000	57.927	-.241	1.083
8	1086.815	60.815	-.253	1.687
9	1161.740	74.926	-.312	1.020
10	1237.063	75.323	-.313	1.239
11	1316.271	79.207	-.330	1.046
12	1396.192	79.921	-.333	1.075
13	1477.199	81.008	-.337	1.076
14	1559.230	82.030	-.341	1.301
15	1644.366	85.136	-.354	1.044

a. The changes are from the previous number of clusters in the table.  
b. The ratios of changes are relative to the change for the two cluster solution.  
c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

表 13-3 聚类结果—各类频数

Cluster Distribution			
Cluster	N	% of Combined	% of Total
1	62	40.8%	39.5%
2	39	25.7%	24.8%
3	51	33.6%	32.5%
Excluded Cases	152	100.0%	96.8%
Total	5		3.2%
	157		100.0%

在一个或几个变量中有缺失值而从分析中剔除。152 个观测量被分配到各类中，62 个分配到第一类，39 个分配到第二类，51 个分配到第三类。

(3) 表 13-4 显示了每类中观测量的均值和标准差, 由于原输出表太宽不好观看, 所以可以把它转动一下。操作方法参考 3.2.2 节中的方法。

转置的结果如表 13-4 所示。它表明了连续变量很好地把各类分开了。1 类中的车辆是便宜的、小(长度、宽度都小)、燃料功效最高。2 类中的车辆特征是适度的价格、较大的汽缸。第三类车辆是昂贵的、大的和适度的燃烧效率。

(4) 表 13-5 是按车辆类型分的频数表, 进一步阐明了各类的特性。第一、三类包括小汽车, 只有第一类中有唯一一款卡车, 查看一下数据文件, 这卡车是 Toyota RAV4。第二类都是卡车。

表 13-4 各类的类中心

Centroids										
			价格(千)	发动机尺寸	马力	轴距	宽度	长度	底盘重量	燃料容量
Cluster 1	Mean		19.61671	2.194	143.24	102.595	68.539	178.235	2.83742	14.979
		Std. Deviation	7.644070	.4238	30.259	4.0799	1.9366	9.6534	.310867	1.8699
	2	Mean	26.56182	3.559	187.92	112.972	72.744	191.110	3.96759	22.064
		Std. Deviation	10.185175	.9358	39.049	9.6537	4.1781	14.4415	.671766	4.2894
	3	Mean	37.29980	3.700	232.98	109.022	72.924	194.688	3.57890	18.443
		Std. Deviation	17.381187	.9493	54.408	5.7644	2.1855	10.3512	.297204	2.0445
Combined	Mean		27.33182	3.049	184.81	107.414	71.089	187.059	3.37618	17.959
	Std. Deviation		14.418669	1.0498	56.823	7.7178	3.4647	13.4712	.636593	3.9376

(5) 图 13-7 是连续变量的并列均值图。这样的图对每个变量生成一个。在聚类中指定了 9 个连续变量共 9 个图。这里只列出两个。图中的每个“工”表现的是各类的均值的 95% 置信区间, 中间的圆点是该类这个变量的均值, 横线是样本中该变量的总均值。

从价格图中可以看出三类的价格置信区间没有交叉, 说明三类的平均价格不同。在价格方面, 三类很好地分开了。发动机尺寸则是第一类较小, 而第二、三类的较大, 这两类的发动机尺寸没有很好地分开。对照表 13-4 也可以看出与其一致的结果。

表 13-5 不同类型车辆的聚类结果频数表

Vehicle type					
		Automobile		Truck	
		Frequency	Percent	Frequency	Percent
Cluster 1		61	54.5%	1	2.5%
2		0	.0%	39	97.5%
3		51	45.5%	0	.0%
Combined		112	100.0%	40	100.0%

(6) 图 13-8 是三类中各连续变量的重要性图。变量以其值递减的顺序放在 y 轴上。横轴是统计量  $t$  值。图中竖线为变量重要性的临界值。对每个要考虑其显著性的变量, 它的  $t$  统计量必须在正或负方向上超过竖线。负  $t$  值表明在该类中, 该变量的值通常比总均值小; 正  $t$  值表明该变量值通常比总均值大。

图 13-8 (a) 是第一类, 图中各连续变量的重要性测度都超过了虚线所示的临界值, 可以得出结论, 所有连续变量对形成第一类都有贡献。燃烧效率比总均值大, 其他变量取值比总均值小。这些结果形成了在表 13-4, 即 Centroids 表中可观测到的趋势。

图 13-8 (b) 是第二类, 看到宽度、长度、马力和价格(千元)对这类的形成不重要。

图 13-8(c) 是第三类, 前后轴间距离、燃料容量对形成这一类不重要, 而燃烧效率刚刚达到有重要意义的程度。



(7) 图 13-9 是分类变量在各类中的重要性图。每类一个图，这里列出一个。它表明分类变量 *type* 在第一类中的重要性。横轴是卡方值，竖线是重要性临界值，超过此线表明该分类变量对该类的形成是重要的。因聚类只指定了一个分类变量，因此只有一条。

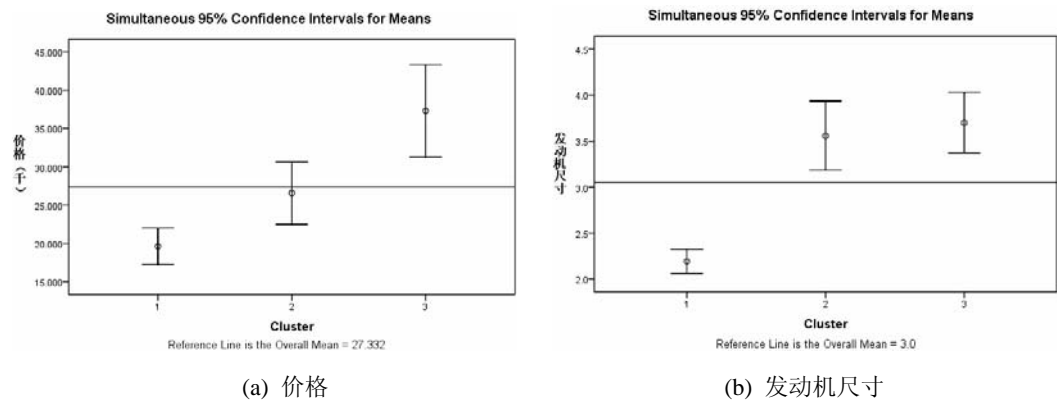


图 13-7 表明连续变量在各类中变异范围的并列均值图

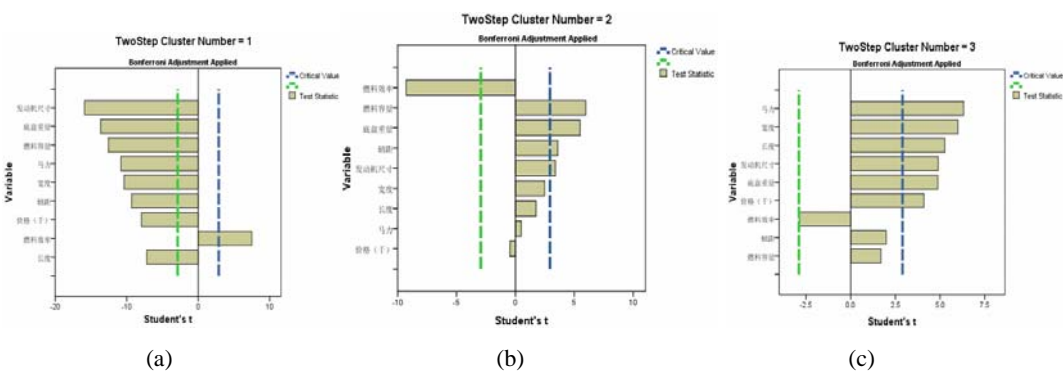


图 13-8 各类中连续变量重要性比较图

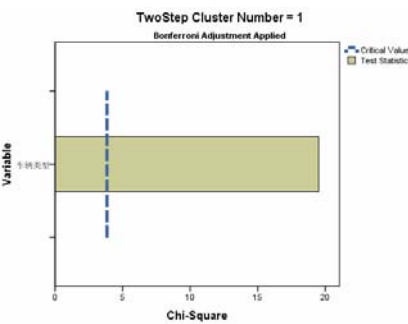


图 13-9 分类变量在第一类中的重要性

使用两步聚类过程 TwoStep Cluster Analysis 已经把车辆分为明显的三类。为了更好地在内部将各类分开，还需要收集有关车辆的其他方面的信息。例如可以注意碰撞试验的成绩或有用的其他项目的信息。

### 13.2.4 两步聚类过程的命令语句

两步聚类过程 TWOSTEP CLUSTER 语句如下：

TWOSTEP CLUSTER

```
[/CATEGORICAL VARIABLES = 分类变量表]
[/CONTINUOUS VARIABLES = 连续变量表]
[/CRITERIA [INITHRESHOLD({0** }{value})] [MXBRANCH({8**}{n })]
[MXLEVEL({3**}{n })]]
[/DISTANCE {EUCLIDEAN }{LIKELIHOOD**}]
[/HANDLENOISE {0**}{n }]
[/INFILE FILE = filename]
[/MEMALLOCATE {64**}{n }]
[/MISSING {EXCLUDE**}{INCLUDE }]
[/NOSTANDARDIZE [VARIABLES = varlist]]
[/NUMCLUSTERS {AUTO** {15**} {n }{FIXED = n } [{AIC }{BIC**}]]
[/OUTFILE [MODEL = filename] [STATE = filename]]
[/PLOT [BARFREQ] [PIEFREQ] [VARCHART [COMPARE {BYCLUSTER**}{BYVAR }]
[NONPARAMETRIC] [CONFIDENCE {95**}{n }] [OMIT] ]]
[/PRINT [IC] [COUNT] [SUMMARY]]
[/SAVE CLUSTER [VARIABLE = varname]]
```

#### 1. TWOSTEP CLUSTER 命令语句

该命令语句调用两步聚类分析过程。有单独的子命令指定参与分析的变量。

#### 2. 子命令

(1) CATEGORICAL VARIABLES 子命令指定参与聚类分析的分类变量。语句关键字后面用等号与分类变量表连接。可以指定若干个分类变量，变量名之间以空格分开。

(2) CONTINUOUS VARIABLES 子命令指定参与聚类分析的数值型连续变量。这是必须要有的语句。语句关键字后面用等号与连续变量表连接。变量表中至少要指定一个连续变量。变量名之间以空格分开。

(3) CRITERIA 语句指定聚类时使用的参数。

① INITHRESHOLD 选项指定初始距离变化极限。在聚类特征树生长过程中，节点间距离小于指定值，CF 树长大同时叶间距离变小、变密；如果大于指定值，就再分出一个节点。系统默认值为 0。

② **MXBRANCH** 选项设置(每个叶子节点的)最大分枝数。一个节点可以具有的最大子节点数。系统默认值是 8, 可以指定一个正整数; 也称作最大树深度。

③ **MXLEVEL** 选项设置 CF 树可以有的最大水平数。系统默认值是 3, 可以指定一个正整数。

(4) **DISTANCE** 子命令指定距离的计算方法。

默认的方法是关键字 **LIKELIHOOD** 指定的负对数似然法; 该方法假定各变量彼此独立, 连续变量服从正态分布, 分类变量服从多项式分布。

如果参与聚类的变量均为连续变量, 也可以使用 **EUCLIDEAN** 关键字指定欧式距离, 即“直线”距离。但是, 如果有非数值变量参与聚类, 指定了欧式距离, 系统会给出错误信息, 并拒绝执行。使用该方法应该注意变量的标准化。

(5) **HANDLENOISE** 子命令指定在聚类过程中, 特征树满了, 既没有叶节点可以接受新观测量进入, 也没有叶节点可以分裂出新节点时, 如何处理待分派的观测量。默认值为 0, 即不进行噪声处理。

如果 CF 树满了, 而 **HANDLENOISE** 值大于 0, 把任意一个数据放到稀松处成为它们自己的噪声叶, CF 树会再长大。如果某处观测量数与最大叶子数之比小于 **HANDLENOISE** 指定的数值, 就被认为是稀少的。在树长大以后, 如果可能, 局外者就被放进 CF 树, 否则在聚类的第二阶段就会被丢弃。

如果树已经满了, 而 **HANDLENOISE** 等于 0, 极限值会被增加, CF 再次长大。聚类结束后, 不能被分派到任意一类的值就被标上是局外者。局外类标识为-1。局外类不包括在聚类数中, 如果指定了  $n$  类, 那么最后聚为  $n$  类和一个局外类也称作噪声类。

(6) **INFILE** 子命令用 **TWOSTEP CLUSTER** 过程修改过去生成的聚类模型, 该模型的 CF 树已经保存在一个用 **OUTFILE** 子命令和 **STATE** 关键字生成的 XML 格式的文件中。该模型将用当前数据文件修改。读者必须使当前工作数据文件中的变量名和顺序与 XML 文件中保存的一致。**TWOSTEP CLUSTER** 仅在内存中修改, XML 文件不变。

如果使用了 **INFILE** 子命令, 如果程序中给出了 **CRITERIA**、**DISTANCE**、**HANDLENOISE** 和 **MEMALLOCATE** 子命令, **TWOSTEP CLUSTER** 会忽略它们。

(7) **MEMALLOCATE** 子命令指定聚类算法所需要的最大内存容量(MB)。如果运算过程中超过了这个最大值, 就使用磁盘去存储内存中装不下的信息。可以指定的最小值是4。如果不使用该语句指定内存空间, 默认的是64MB。你的系统允许指定的最大值是多少, 可以请示你的系统管理员。

(8) **MISSING** 子命令指定怎样处理读者定义的缺失值的观测量。默认**EXCLUDE**。**TWOSTEP CLUSTER**删除带有系统缺失值的观测量。

**EXCLUDE** 指定分析中剔除带有系统缺失值或读者缺失值的观测量。是默认的选择。

**INCLUDE** 指定读者缺失值作为有效值处理, 分析中剔除带有系统缺失值的观测量。

(9) **NOSTANDARDIZE** 子命令指定不进行标准化的连续变量。如果没有使用这个子

命令, **TWOSTEP CLUSTER** 对所有参与分析的连续变量进行标准化。每个变量都减去该变量的均值除以标准差。如果程序中使用了 **NOSTANDARDIZE** 子命令, 但没有变量表, **TWOSTEP CLUSTER** 不会对任何连续变量进行标准化。

(10) **NUMCLUSTERS** 子命令指定聚类最后结果要求的类数。

**AUTO** 自动选择类数, 可以指定一个最大可能的类数。两步聚类在 1 到指定数之间根据指定的判据寻找最好的聚类结果。确定类数的判据可以通过选项 **BIC** 或 **AIC** 指定。

**FIXED** 在等号后面给出一个正整数, 作为最后的类数。

(11) **OUTFILE** 子命令指定输出文件的内容、存储路径和文件名。输出文件格式均为 XML 文件。

**MODEL** 关键字后面用等号连接保存聚类模型的路径和文件名。

**STATE** 关键字后面用等号连接保存聚类特征树的路径和文件名。

(12) **PLOT** 子命令

① **BARFREQ** 输出表明每类中观测量数的条形图。

② **PIEFREQ** 输出表明每类中观测量数和占总数百分比的饼图。

③ **VARCHART** 指定对每类输出表明该类中变量重要性的条形图。变量按重要性大小排序。该选项中, 还有以下几个子选项:

- **COMPARE** 要求对重要性进行比较, **BYCLUSTER** 是默认的选项, 要求比较变量在各类中的重要性, 对每变量输出一个重要性图; **BYVAR** 要求在每类中比较各变量的重要性, 对每类输出一个重要性图;

- **NONPARAMETRIC** 该关键字告诉两步聚类过程在图中用什么测度变量重要性, 如果没给出这个关键字, 将对分类变量给出 **Pearson** 卡方, 对连续变量给出  $t$  统计量。如果给出了这个选项, 重要性是一个针对均值相等的检验的  $1-p$  值表示。

- **CONFIDENCE** 对重要性设置置信水平。该关键字使得两步聚类过程把 1 减去该关键字设置值认为是类内变量分布与整体分布相等的检验的  $\alpha$  值。可以指定 1~100 之间任何值。如果给出了 **NONPARAMETRIC** 或 **BYVARIABLE** 关键字, **CONFIDENCE** 指定的值在变量重要性图上只会表示为一根垂直线。默认值为 95。

- **OMIT** 忽略不重要的变量。在图中不显示那些重要性在  $\alpha$  (即  $1 - \text{CONFIDENCE}$ ) 水平上不显著的变量。

(13) **PRINT** 子命令要求两步聚类过程输出与各类有关的表格。内容有:

① **IC** 选项要求输出聚类过程中的 **AIC** 或 **BIC** 的值。选择了哪个, 就输出哪个的值。如果在 **NUMCLUSTERS** 子命令中, 没有使用 **AUTO** 选项, 两步聚类将忽略 **IC** 选项输出判据的要求, 给出警告信息。如果使用了 **AUTO** 选项, 但是, 后面跟着的最大类数是 1, 也会忽略 **IC** 选项的要求。

② **SUMMARY** 选项输出两个表格。一个表格分类给出连续变量的均值、标准差。另一个表格分类给出分类变量各水平的观测量数。

③ COUNT 选项输出各类中的观测量计数。

(14) SAVE 子命令要求生成并保存新变量到工作数据文件。

CLUSTER 要求生成并保存类标号变量。可以使用选项 VARIABLE, 在等号后面指定变量名。如果不用这个选项, 系统给出默认的变量名是 TSC<sub>*n*</sub>, 这里的 *n* 表示在同一个 SPSS 运行期间, 产生的表示分类标号的新变量的顺序。

AIM 过程语句如下:

```
AIM 分组变量 [/CATEGORICAL 分类变量] [/CONTINUOUS 变量表]
[/CRITERIA [ADJUST = {BONFERRONI**}] [CI = {95**}]
{NONE} {value}
[HIDENOTSIG = {NO**}] [SHOWREFLINE = {NO}] ]
{YES} {YES**} [/MISSING {EXCLUDE**}] {INCLUDE}
[/PLOT [CATEGORY] [CLUSTER [(TYPE = {BAR*})]] [ERRORBAR]{PIE}
[IMPORTANCE [(X = {GROUP*}] [Y = {TEST*}]])] {VARIABLE} {PVALUE}
```

AIM 过程是一个补充的与 TWOSTEP CLUSTER 配合使用的作图的过程语句。有的选项随着 TWOSTEP CLUSTER 子命令中的选项改变, 例如置信水平 CI 值与 CONFIDENCE 设置的值相同。基本语句有:

(1) CATEGORICAL 子命令指定参与分析的分类变量。

(2) CONTINUOUS 子命令指定参与聚类分析并在统计图中出现的连续变量。

以上两个子命令与 TWOSTEP CLUSTER 中相应语句指定的变量名必须一致。

(3) PLOT 子命令指定生成的统计图。

① ERRORBAR 生成各类的变量均值置信区间对比图。

② CATEGORY CLUSTER (TYPE=PIE) 生成饼图。

③ IMPORTANCE 关键字指定生成变量重要性图。

④ CRITERIA ADJUST 子命令指定与聚类结果和统计图有关的参数。在 ADJUST 关键字后的等号后面指定使用的判据, 贝叶斯判据关键字 BONFERRONI; CI 在其后的等号后面指定置信水平, 默认为 95; SHOWREFLINE 关键字指定是否显示参考线。等号后面可以是 YES 或 NO; HIDENOTSIG 指定重要性不显著的变量是否在图中出现。等号后面 NO 表示不出现, YES 表示要在图中出现。

## 13.3 快速样本聚类

### 13.3.1 快速样本聚类概述

当要聚成的类数确定时, 使用 Quick Cluster 过程可以很快将观测量分到各类中去。其特点是处理速度快, 占用内存少, 适用于大样本的聚类分析。

**K-Means Cluster** 执行 **Quick Cluster** 命令, 使用  $k$  均值分类法对观测量进行聚类, 可以完全使用系统默认值执行该命令, 也可以对聚类过程设置各种参数进行人为的干预。例如可以事先指定把数据文件中的观测量分为几类, 指定使聚类过程中止的收敛判据或迭代次数, 指定聚类结束后在的输出窗口中显示哪些内容, 是否将聚类结果或中间数据存入输出数据文件, 指定其文件名以及把哪些数据存入数据文件等。

进行快速样本聚类首先要选择用于聚类分析的变量和类数。参与聚类分析的变量必须是数值型变量, 且至少要有一个。为了清楚地表明各观测量最后聚到哪一类, 还应该指定一个表明观测量特征的变量作为标识变量, 例如编号、姓名之类的变量。聚类数必须大于等于 2, 但类数不能大于数据文件中的观测量数。

如果选择了  $n$  个数值型变量参与聚类分析, 最后要求聚类数为  $k$ , 那么可以由系统首先选择  $k$  个观测量 (也可以由读者指定) 作为聚类的种子,  $n$  个变量组成  $n$  维空间。每个观测量在  $n$  维空间中是一个点。  $k$  个事先选定的观测量就是  $k$  个聚类中心点, 也称为初始类中心。按照离这几个类中心的距离最小原则把观测量分派到各类中心所在的类中去, 构成第一次迭代形成的  $k$  类。根据组成每一类的观测量, 计算各变量均值。每一类中的  $n$  个均值在  $n$  维空间中又形成  $k$  个点。这就是第二次迭代的类中心。按照这种方法依次迭代下去, 直到达到指定的迭代次数或达到中止迭代的判据要求时, 迭代停止, 聚类过程结束。

快速聚类使用的是欧氏距离平方, 各变量权数相等。如果使用其他统计量进行聚类, 必须使用 **Hierarchical Cluster** 进行聚类分析。快速聚类变量必须是连续变量。如果测定变量值的单位不同, 应该对聚类变量使用 **DESCRIPTIVES** 过程进行标准化后再进行聚类分析, 否则会得出错误的结论。如果聚类变量是计数变量或二值变量, 则使用 **Hierarchical Cluster** 分析过程进行聚类分析。

### 13.3.2 快速样本聚类过程

快速聚类过程适用于对大样本进行快速聚类, 尤其是对形成的类的特征 (各变量值范围) 有了一定认识时, 此方法使用起来会更加得心应手。操作方法分为以下几步:

1. 建立或读入数据文件后, 按 **Analyze**→**Classify**→**K-Means Cluster** 顺序单击鼠标, 展开如图 13-10 所示的对话框。
2. 指定分析变量和标识变量。在源变量表中选择参与聚类分析的数值型变量, 移到右面的 **Variables** 框中; 选择能唯一标识各观测量的变量, 送入 **Label Cases By** 栏中。
3. 确定分类数。在 **Number of Clusters** 框中显示系统默认分为两类。可按分析要求输入分类数。
4. 选择聚类方法。在 **Method** 栏中的两项中可以选择一种聚类方法。

(1) 系统默认 **Iterate and classify**, 聚类的迭代过程中使用 **K-Means** 算法不断计算类中心, 并根据结果更换类中心, 把观测量分派到与之最近的以类中心为标志的类中去。

(2) **Classify Only**, 根据初始类中心进行聚类。在聚类过程中不改变类中心。

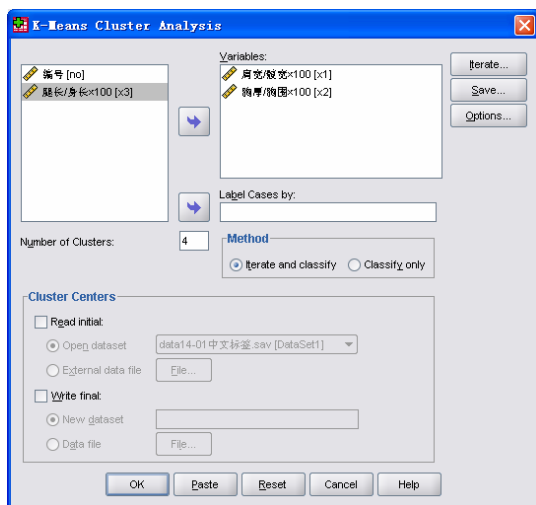


图 13-10 K-Means Cluster Analysis 主对话框

**External data file** 下 **File** 按钮后面显示包括路径的文件全名。

选择 **Read initial**, 需要事先建立一个数据集, 其中观测量的数目与要聚成的类数相等, 每个观测值都由参与聚类的变量值组成。

(2) **Write final**, 要求把聚类结果中的各类中心数据保存到指定的文件中, 该文件可以作为以后聚类的初始类中心文件。

① 选择 **New dataset** 在后面输入数据文件名, 运行结果会把最后结果的类中心保存在指定的文件中。注意这个文件无需指定保存位置, 所以结束 SPSS 前要保存这个文件, 否则会丢失。

② 选择 **Data file**, 单击 **File** 按钮, 在 **Write to File** 对话框中指定文件保存位置 (路径) 和文件名。按 **Save** 按钮返回。在 **File** 按钮后面显示包括路径的文件全名。

6. 输出数据选项。在主对话框中单击 **Save** 按钮, 展开保存新变量对话框, 见图 13-11。

(1) 选择 **Cluster membership**, 要求在当前工作数据文件中 (数据窗口中) 建立一个新变量, 默认变量名为 **qcl\_1**。其值表示聚类结果, 为类顺序标号 1, 2, 3, ..., 即表明各观测值被分配到哪一类。

(2) 选择 **Distance from cluster center**, 要求在当前数据窗口中建立一个新变量, 默认变量名为 **qcl\_2**。变量值为各观测值距所属类的类中心间的欧氏距离。

(1) 在 **Maximum Iterations** 参数框中限定 K-Means 算法中的迭代次数。当达到限定

5. 在主对话框中 **Cluster Centers** 栏内选择初始类中心。

(1) **Read initial**, 要求使用指定数据文件中的观测值作为初始类中心。选择此项后, 还需要选择:

① **Open dataset** 如果包含种子观测值的数据文件已经打开, 单击向下箭头, 在下拉菜单中选择一个其观测值作为初始类中心的数据集。

② **External data file** 如果包含种子观测值的数据文件没有打开, 单击 **File** 按钮, 在 **Read from File** 对话框中指定文件所在位置 (路径) 和文件名。该文件的观测值作为初始类中心的数据。按 **Open** 按钮返回。在



图 13-11 保存新变量对话框

的迭代次数时,即使没有满足收敛判据,迭代也停止。系统默认值为迭代 10 次,选择范围为 1~999。

(2) 在 **Convergence Criterion** 参数框指定 **K-Means** 算法中的收敛判据,其值必须大于等于 0,小于 1,默认值为 0。该项数值等于  $N$  的含义是,当两次迭代计算的最小的类中心的变化距离小于初始类中心距离的  $N\% \times 100$  时迭代停止。例如判据设置为 0.02,当一次完整的迭代不能使任何一个类中心距离的移动(变化量)与原始类中心距离的比小于 2% 时,迭代停止。

(3) 若设置了以上两个参数,在迭代过程中,满足了其中一个参数,迭代就停止。

(4) **Use running means**,限定在每个观测量被分配到一类后即刻计算新的类中心。如果不选择此项,在完成了所有观测量的一次分配后再计算各类的类中心,这样节省迭代时间。

#### 7. 控制聚类分析过程的选项。

在主对话框中单击 **Iterate** 按钮,展开设置迭代参数的对话框,如图 13-12 所示。只有在 **Method** 栏中选择了 **Iterate and classify** 项,才激活此项,打开此对话框对迭代次数和聚类判据进行进一步选择。

8. 在主对话框中单击 **Options** 按钮,打开如图 13-13 的选择项对话框。在此指定要计算的统计量和对带有缺失值的观测量的处理方式。

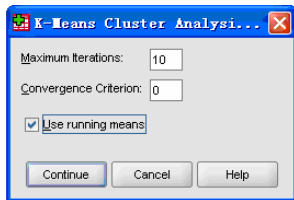


图 13-12 迭代参数对话框

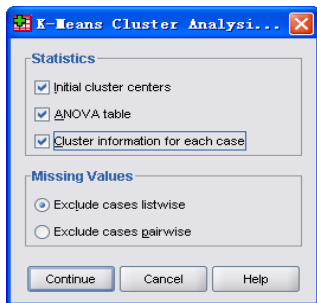


图 13-13 选择项对话框

(1) 在 **Statistics** 栏中选择要计算和输出的统计量:

① **Initial cluster centers**, 初始类中心。

② **ANOVA table**, 方差分析表。

③ **Cluster information for each case**, 每个观测量的分类信息,如最终所属类和该观测量距所属类中心的距离。

(2) 在 **Missing Values** 栏中选择一种处理带有缺失值观测量的方法。

① **Exclude cases listwise**, 从分析中剔除在 **Variables** 表中变量带有缺失值的观测量。

② **Exclude cases pairwise**, 只有当一个观测量的全部聚类变量值均缺失时才将其从分析中剔除,否则根据所有其他非缺失变量值,把它分配到最近的一类中去。



13.3.3 快速样本聚类分析实例

【例 2】本节对游泳运动员进行聚类，以便分项。为简化问题，仅以 10 名运动员的三项测试数据为例。其中变量为指标 1：x1（肩宽/髋宽×100），指标 2：x2（胸厚/胸围×100），指标 3：x3（腿长/身长×100），预计按姿势分为蝶泳、仰泳、蛙泳、自由泳四类。打开数据编号为 data13-02 的数据文件，操作步骤如下：

- (1)按 Analyze→Classify→K-Means cluster 顺序单击菜单项，打开主对话框。
- (2) 此例题要求根据 x1~x3 进行聚类，因此选择这三个变量，移至 Variables 矩形框中。选择变量 no 作为标识变量送入 Label Cases By 框中。
- (3) 泳姿分四类，在 Number of Cluster 后面的编辑栏中输入数字 4。
- (4) 其他使用系统默认参数。提交运行的程序如下：

```
QUICK CLUSTER      x1 x2 x3                                ①
/MISSING=LISTWISE                                         ②
/CRITERIA= CLUSTER(4) MXITER(10) CONVERGE(0)              ③
/METHOD=KMEANS(NOUPDATE)                                ④
/PRINT ID(no ) INITIAL.                                    ⑤
```

语句解释如下：

- ① 调用快速聚类过程，使用 x1，x2，x3 对观测量进行聚类分析。
  - ② 要求分析中剔除分析变量中任一变量有缺失值的观测量。
  - ③ 要求迭代结果将观测量分为 4 类，最大迭代次数 10，收敛判据为 0。即迭代到 10 次或收敛到 0 时，迭代停止。
  - ④ 要求使用 k 均值法进行聚类，但聚类过程中不改变类中心。
  - ⑤ 指定按标识变量为变量 no 打印聚类结果和初始类中心。
- (5) 输出结果见表 13-6~表 13-8。

表 13-6 初始类中心

	Initial Cluster Centers			
	Cluster			
	1	2	3	4
肩宽/髋宽×100	125	122	120	120
胸厚/胸围×100	20	18	17	19
腿长/身长×100	44	43	42	44

表 13-7 两次迭代后类中心的变化

Iteration	Iteration History <sup>a</sup>			
	Change in Cluster Centers			
	1	2	3	4
1	.707	.354	.707	.707
2	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 2.449.

表 13-6 初始类中心，由于读者没有指定聚类的初始类中心，此表中的作为类中心的观测量是由系统确定的。表中给出作为 4 类初始类中心的观测量各变量值。

表 13-7 所示是两次迭代后，类中心的变化。由于没有指定迭代次数或收敛判据，因此使用系统默认值：最大迭代次数为 10，收敛判据为 0。本快速聚类过程执行 2 次迭代后，类中心的变化为 0，迭代就停止了。表 13-7 给出了每次迭代类中心的变化量。

表 13-8 给出了聚类结果形成的四类的类中心的三个变量的值。右表显示的是聚类结

果, 每类中观测量的数目。除第二类有 4 个外, 其余各类, 均有 2 个运动员。

由上述结果输出可以看出, 采用系统默认值的输出结果并不令人满意。若想知道某个观测量属于哪一类, 从输出信息中找不到。因此需要使用选项。

表 13-8 最终的四类的类中心和聚类总结

Final Cluster Centers					Number of Cases in each Cluster	
	Cluster				Cluster	
	1	2	3	4		
肩宽/腋宽×100	124	122	120	120	1	2.000
胸厚/胸围×100	20	18	17	19	2	4.000
腿长/身长×100	44	43	42	44	3	2.000
					4	2.000
					Valid	10.000
					Missing	.000

### 【例 3】指定初始类中心的聚类方法例题

数据仍为 data13-02。已知  $no=9、8、4、6$  的四名运动员分别是蝶、仰、蛙、自由四种姿势成绩突出者, 以这 4 个观测量作为初始聚类中心进行聚类。操作步骤如下:

(1) 建立包含初始聚类中心 4 个观测量的数据文件, 类中心数据文件 (也称种子数数据文件) data13-02a 存入磁盘中。要求种子数据文件:

- ① 其格式必须与 data03-02 中数据文件格式相同。
- ② 文件中的变量必须在当前工作数据文件中存在, 并且变量名相同, 在即将进行的快速聚类中也选择相同的变量作为聚类变量。
- ③ 其中的观测量数必须与在主对话框中指定的类数相同。
- ④ 有一个表明类号的变量, 变量名为 cluster\_。

该数据文件可以是前一次快速聚类产生的输出文件, 也可以根据经验找出的最具代表性的观测量作为初始类中心, 或称聚类的种子。

(2) 将已经存在的原始数据 data13-02 调入, 显示在当前数据窗口中。打开作为种子的数据文件 data13-02a。

(3) 首先, 按前面例题中的(1)~(3)步选择聚类变量、标识变量, 指定分类数。

(4) 在 Cluster Center 栏内选中 Read initial 组, 选择 Open dataset, 其后的框中显示 data13-02a.sav[datasetn], 即已经打开的数据文件名指定为初始类中心文件。

(5) 选中 Write final, 选择 Data file 保存聚类结果的类中心, 按 File 按钮, 指定存储作为以后种子观测量的数据文件。指定存取路径和文件名 (data13-02b)。

(6) 聚类方法选择, 在主对话框 method 栏中选择 Iterate and classify 项。

(7) 聚类过程控制参数仍选用系统默认值。Iterate 项内的值保持不变。

(8) 单击 Save 按钮, 在对话框中选择 Cluster membership 和 Distance from cluster center 复选项。

(9) 单击 Options 按钮, 打开相应的对话框, 选中 Statistics 栏中的全部复选项。由于数据文件中没有缺失值, 故保持 Missing Value 组中系统默认的处理方式。

(10) 主对话框中单击 OK 按钮提交系统执行。执行的程序如下:

```

QUICK CLUSTER x1 x2 x3                                ①
/MISSING=LISTWISE                                     ②
/CRITERIA= CLUSTER(4) MXITER(10) CONVERGE(0)         ③

```

```
/METHOD=KMEANS(NOUPDATE)
/SAVE CLUSTER DISTANCE
/PRINT ID(no ) INITIAL ANOVA CLUSTER DISTAN
/FILE='D:\data13-02a.SAV'
/OUTFILE=' D:\data13-02b.SAV'.
```

④  
⑤  
⑥  
⑦  
⑧

以上程序语句的①②③④语句与前一例题的相应语句相同。

第⑤子命令语句要求生成新变量，一个新变量的值是各观测量所属的类号；另一变量的值是该观测量与所属类的类中心之间的距离。

第⑥子命令语句要求输出：**ID** 即各类成员的标识变量的值，本例的标识变量是运动员编号 *no*；**INITIAL** 初始类中心；**ANOVA** 方差分析表；**CLUSTER DISTAN** 各类之间距离（类中心间的距离）。

第⑦子命令语句指定了聚类种子文件。可以改为：

`/initial (122,17,42 122,19,43 124,20,45 120,19,44)` 即把四个初始聚类中心观测量的 *x1*、*x2*、*x3* 值列在 **Initial** 子命令后面括号中，与 **File** 子命令是等价的。

第⑧子命令语句指定了聚类结果的输出文件的路径和文件名。

(11) 输出结果见表 13-9～表 13-15。数据文件中的聚类种子数据见图 13-14。

(12) 结果解释

表 13-9 中的初始类中心与 **File** 子命令中指定的种子文件 *data13-02a* 中的数据一致。

表 13-10 表明共经过两次迭代完成聚类。第一次迭代 1~4 类的类中心与初始类中心之间的距离分别为 0.707、0.707、0.745、1.054。从操作过程和程序语句的 `criteria=cluster(4) Maxiter(10)Converge(0)`子命令看到结束聚类过程的判据有两个，一个是最大迭代次数为 10，一个是类中心变化距离为 0。从表 13-5 看到当进行了第二次迭代后，类中心几乎没有变化，使用判据 0，结束了聚类过程。

表 13-9 初始类中心（根据子命令 Initial）

	Initial Cluster Centers			
	Cluster			
	1	2	3	4
肩宽/臂宽×100	124	120	122	122
胸厚/胸围×100	20	19	19	17
腿长/身长×100	45	44	43	42

Input from FILE Subcommand

表 13-10 迭代过程中类中心的变化量

Iteration	Change in Cluster Centers			
	1	2	3	4
1	.707	.707	.745	1.054
2	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 2.236.

表 13-11 给出了聚类结果，即每个观测量用 *no* 标识，表头的“编号”为变量 *no* 的标签。**Cluster** 的值为类号，表明各观测量最终被分配到哪一类，**Distance** 的值为该观测量在三维坐标中的点与类中心点的距离。如果选择的类中心是各类最具代表性的观测量，则 **Distance** 值越大，与该类代表性观测的差异越大。

表 13-12 给出了 4 个类中心的 3 个变量值，即类中心在三维坐标空间中的位置。

表 13-13 给出的是聚类结束时, 两两类中心间的距离。表格第一行和左边第一列均为类号。两类间的距离在行、列交叉点单元格中。

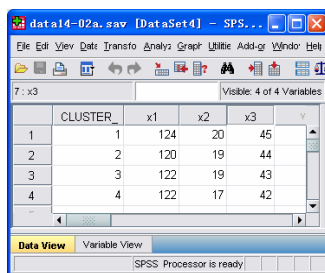
表 13-14 是方差分析表。3 个变量中任意一个变量的类间均方 (Cluster MS) 都远远大于类内的误差均方值 (Error MS)。从概率值来看, 3 个变量使类间无差异的假设成立的概率均小于 0.1%。方差分析结果表明, 参与聚类分析的 3 个变量能很好地区分各类, 类间的差异足够大。聚类的方差分析检验的零假设应该是: 类均值相等 (各类间无差异)。该分析结果可用于描述分类的目的。

表 13-15 是聚类总结。Cluster 给出了各类的观测量数、参与分析的合法观测量数 Valid 和缺失值数 Missing。可以看出指定了初始类中心 (种子) 的结果与没使用初始类中心的结果略有不同 (与表 13-8 比较)。

在 Notes 中给出了生成的新变量的信息。如果输出信息区没有显示 Notes, 在导航器中单击 Notes 图标可以打开并看到 Notes 的内容。

表13-11 各观测量所属类成员表

Cluster Membership			
Case Number	编号	Cluster	Distance
1	1	1	.707
2	2	3	.745
3	3	4	1.054
4	4	1	.707
5	5	3	.471
6	6	2	.707
7	7	4	.667
8	8	3	.745
9	9	4	1.054
10	10	2	.707



CLUSTER_	x1	x2	x3	y
1	124	20	45	
2	120	19	44	
3	122	19	43	
4	122	17	42	

图 13-14 聚类种子数据

表 13-12 最终的类中心的变量值

	Cluster			
	1	2	3	4
肩宽/髋宽×100	124	120	122	121
胸厚/胸围×100	20	19	18	17
腿长/身长×100	44	44	43	42

表 13-13 最终的类中心间的距离

Cluster	Distances between Final Cluster Centers			
	1	2	3	4
1		4.123	3.613	5.411
2	4.123		2.014	3.504
3	3.613	2.014		2.000
4	5.411	3.504	2.000	

表 13-14 方差分析

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
肩宽/髋宽×100	6.644	3	.611	6	10.873	.008
胸厚/胸围×100	3.911	3	.111	6	35.200	.000
腿长/身长×100	4.644	3	.278	6	16.720	.003

The F tests should be used only for descriptive purposes because the clusters have been c maximize the differences among cases in different clusters. The observed significance leve corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster r equal.

表 13-15 聚类总结

Number of Cases in each Cluster		
Cluster	1	2.000
	2	2.000
	3	3.000
	4	3.000
Valid		10.000
Missing		.000

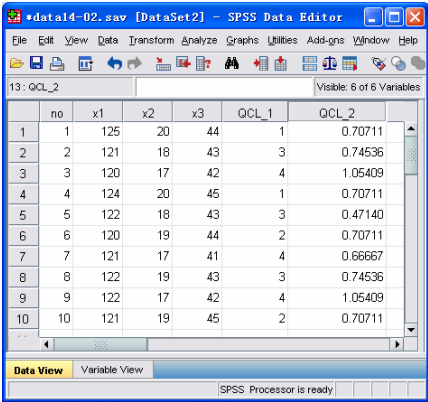


图 13-15 工作数据文件中的新变量

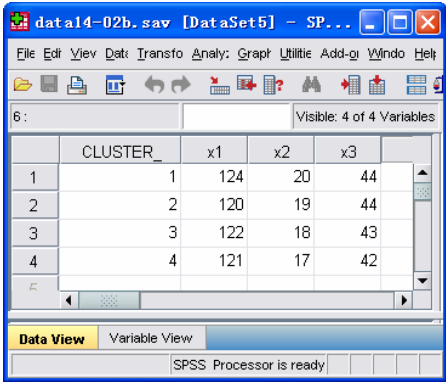


图 13-16 输出数据文件

图 13-15 是当前工作数据文件。根据指定的选项共建立了两个新变量，显示在工作数据文件窗口中。QCL1 是类号，QCL2 是观测量距所属类的类中心之间的距离。

输出数据文件中只有最终的类中心数据，见图 13-16。此输出数据文件可以作为对另一个样本进行快速聚类的初始类中心了。

### 13.3.4 快速样本聚类过程的命令语句

#### 1. 命令格式

```
QUICK CLUSTER {varlist}{ALL }  
    [/MISSING={LISTWISE**} {PAIRWISE } {DEFAULT }} [INCLUDE]] [/FILE=file]  
    [/INITIAL=(value list)]  
    [/CRITERIA=[CLUSTER({2**}{n})][NOINITIAL][MXITER({10**}{n})]  
    [CONVERGE({0**}{n } )]]  
    [/METHOD=[{KMEANS([NOUPDATE])**} {KMEANS(UPDATE)} {CLASSIFY }]  
    [/PRINT=[INITIAL**] [CLUSTER] [ID(varname)] [DISTANCE] [ANOVA] [NONE]]  
    [/OUTFILE=file] [/SAVE=[CLUSTER[(varname)]] [DISTANCE[(varname)]]]
```

其中：QUICK CLUSTER 是命令关键字，使用 K-Means 算法对数据文件中的观测量进行快速聚类。该语句必须是第一个语句，其余子命令的顺序任意。

变量表和 ALL 任选其一，只有当数据文件中的变量全部是数值型变量，而且全部作为分析变量时才可以选用 ALL。否则应该把参与分析的变量名称一一列出，用空格隔开。所有子命令均为可选择的子命令。

#### 2. 子命令及其含义

(1) MISSING 子命令指定处理带有缺失值的观测量的方式，选项有三个：

- ① LISTWISE 是默认项，如果观测量的分析变量值为缺失值，分析时剔除该观测量。
- ② PAIRWISE 对分析变量不全为缺失值的观测量，根据其非缺失变量值分派到与之

最近的类中去。只有当分析变量值全部为缺失值的观测量，才在分析时剔除。

③ **DEFAULT** 与 **LISTWISE** 相同。

④ **INCLUDE** 把读者定义的缺失值当作合法值处理。

前三个选项只能选其一，但都可以与 **INCLUDE** 同时使用。

(2) **FILE** 子命令和 **INITIAL** 子命令都用于指定初始类中心。

① **FILE** 子命令指定包含初始类中心的文件名和存储位置。文件名用单引号引起来。

例如：`/FILE='D:\SPSSSAV\data13-02a.SAV'` 是一个符合语法要求的 **FILE** 子命令。注意引号必须是英文半角。

② **INITIAL** 子命令直接指出作为初始类中心的观测量，列出各类中心的分析变量值。变量值列出的顺序，应该与主命令变量表的顺序一致。各类中心变量表之间，用空格分开。例如：`INITIAL(122,17,42 122,19,43 124,20,45 120,19,44)` 是合法的 **INITIAL** 子命令。

如果主命令为 **QUICK CLUSTER x1 x2 x3**，上述 **INITIAL** 告诉系统四个初始类中心的三个变量值分别是：第一类中心： $x_1=122$ ， $x_2=17$ ， $x_3=42$ ；第二类中心： $x_1=122$ ， $x_2=19$ ， $x_3=43$ ；第三类中心： $x_1=124$ ， $x_2=20$ ， $x_3=45$ ；第四类中心： $x_1=120$ ， $x_2=19$ ， $x_3=44$ 。

注意：与此相配合的应该有子命令：**CRITERIA=CLUSTER(4)**。

(3) **CRITERIA** 子命令为快速聚类提出聚类数目。有以下 4 个选项：

① **CLUSTER(*k*)** *k* 是正整数，指定最后聚成 *k* 类；**CLUSTER(2)** 是系统默认值。

② **NOINITIAL** 不指定初始类中心。

③ **MXITER(*n*)** 指定迭代次数，*n* 为 1~999 间的整数。系统默认 **MXITER(10)**。

④ **CONVERGE(*n*)** *n* 是正整数， $1 > n \geq 0$ ，聚类停止的收敛判据。当两次迭代间，类中心变化的距离小于初始类中心最小距离乘以 *n* 的积时，迭代停止。0 是系统默认值。

(4) **METHOD** 是选择聚类方法的子命令。选项有两个：

① **K-MEANS** 法，即迭代法。根据何时重新计算新类中心又分为：

`/METHOD=KMEANS(UPDATE)` 每分派一个观测量到某类，就重新计算一次该类的新类中心。

`/METHOD=KMEANS(NOUPDATE)` 所有变量都分派到各类，一次迭代完成再计算一次各类的新类中心。

② **CLASSIFY** 法，指定初始类中心，不进行迭代，按最近原则把观测量分派到各类中心的类中去。子命令为：`/METHOD=CLASSIFY`。

(5) **PRINT** 子命令指定在输出窗口中显示的结果数据。共有 6 个选项：

① **INITIAL** 显示初始聚类中心。

② **CLUSTER** 显示各观测量所属的类。

③ **ID** (标识变量名) 显示各观测量所属类的同时，显示标识变量值。

④ **DISTANCE** 显示各观测量到所属类中心的距离。

⑤ ANOVA 显示方差分析表。

⑥ NONE 不输出任何信息。

(6) OUTFILE 子命令定义输出数据文件的名称, 包括其存储位置。存储路径和文件名, 用英文半角引号括起来。聚类过程结束后, 将各类的类中心存入该数据文件。可以利用该文件配合/METHOD=CLASSIFY 子命令, 对未知观测量数据集进行分类。

(7) SAVE 子命令在输入数据文件中建立新变量, 存入聚类结果数据。并为新变量指定变量名。在选项中把新变量名写在括号中。选项有两个:

① CLUSTER (变量名) 读者在括号中指定新变量名, 其值为各观测量所属的类。默认的变量名为 qcl\_1。

② DISTANCE (变量名) 读者在括号中指定新变量名, 其值为各观测量与类中心之间的距离。默认的变量名为 qcl\_2。

## 13.4 分层聚类

### 13.4.1 分层聚类概述

#### 1. 分层聚类概述

聚类的方法有多种, 除了前面介绍的两步聚类和快速聚类法外, 最常用的是分层聚类法。根据聚类过程不同又分为分解法和凝聚法。

(1) 分解法: 聚类开始时把所有个体 (观测量或变量) 都视为属于一大类, 然后根据距离和相似性逐层分解, 直到参与聚类的每个个体自成一类为止。

(2) 凝聚法: 聚类开始时把参与聚类的每个个体 (观测量或变量) 视为一类, 根据两类之间的距离或相似性逐步合并, 直到合并为一个大类为止。

无论哪种方法其聚类原则都是相近的聚为一类, 即距离最近或最相似的聚为一类。实际上以上两种方法是方向相反的两种聚类过程。

#### 2. Cluster 过程的功能

分层聚类的方法可以用于样本聚类 (Q 型), 也可以用于变量聚类 (R 型)。通常情况下在聚类进行之前, PROXIMITIES 过程先根据反映各类特性的变量对原始数据进行预处理, 即利用标准化方法对原始数据进行一次转换, 并计算相似性测度或距离测度。然后 Cluster 过程根据转换后的数据进行聚类分析。SPSS 的分层聚类的各方法都包含了 PROXIMITIES 过程对数据的处理, CLUSTER 过程对数据聚类。给出的统计量可以帮助读者确定最好的分类结果。

Cluster 过程可以通过 Plot 选项, 给出两种图: Dendrogram 树形图和 Icicle 冰柱图。

Cluster 过程的输出项可以选择, 还可以建立新变量, 把聚类结果, 即每个个体被分派到的类编号作为新变量的值, 保存到当前的工作数据文件中。

### 3. 在 Cluster 过程中使用的术语

(1) 聚类方法。实现分层聚类的具体方法有许多种，各种方法的区别在于如何定义和计算两项（两个个体、两类或个体与类）之间的距离或相似性。这一点体现在聚类方法（Method）的一系列选项上。如果不熟悉对聚类方法的定义，可以使用系统默认的方法。需要确定的选项有：

聚类法的选择：定义计算两项间距离和相似性的方法，默认使用组间平均连接法。

测度方法的选择：对距离和相似性的测度方法又有多种，这一点体现在测度方法的选择上。如果对测度方法不熟悉，可以采用系统默认的欧氏距离平方。

定义距离和相似性的方法不同，测度距离和相似性的算法就不同，会导致聚类结果稍有区别，但大体上是一致的。

(2) 标准化。如果参与聚类的变量的量纲不同会导致错误的聚类结果。因此在聚类过程进行之前必须对变量值进行标准化，即消除量纲的影响。用不同方法进行标准化，会导致不同的聚类结果，因此在选择标准化方法时要注意变量的分布。如果是正态分布应该采用 Z 分数法。如果参与聚类的变量量纲相同，可以选择默认值 None，对数据不进行标准化处理。

(3) 树形图。表明每一步中被合并的类及系数值，把各类间的距离转换成 1~25 间的数值。

(4) 冰柱图。把聚类信息综合到一张图上。如果作纵向冰柱图，则参与聚类的个体各占一行，标以个体（观测量或变量）号或在图纸允许的情况下标以个体的标签；聚类过程中的每一步占一行，标以步的顺序号。如果作横向冰柱图，则参与聚类的个体（观测量或变量）各占一行，聚类的每一步占一行。如果不指定加以限制的选项，则显示聚类的全过程。

树形图和冰柱图都是最后确定分类结果的重要手段。因为无论凝聚法还是分解法均不给出确定的分类结果。最后的分类结果需要读者根据研究的对象和研究目的自己确定。

## 13.4.2 分层聚类过程

无论是观测量聚类还是变量聚类均按下述步骤进行。我们在叙述操作步骤的过程中对涉及的选项及其含义分别加以说明。

1. 在数据窗口中建立工作数据文件。

2. 按 Analyze→Classify→Hierarchical Cluster 顺序单击菜单项，展开如图 13-17 所示的分层聚类分析主对话框。

3. 在对话框中部的 Cluster 栏中选择聚类类型：

(1) 选择 Variables 项，要进行变量（R 型）聚类。

(2) 选择 Cases 项，要进行观测量（Q 型）聚类。

4. 指定参与分析的变量，即能反映分类特征的变量，送入 Variable(s)框中。如果做



观测量聚类，还要选择能唯一标识观测量的变量，移到右侧的 Label Cases by 框中。

5. 如果参与分析的变量量纲一致，不必对数据进行标准化，其余的全部选择系统默认值，此时就可以单击 OK 按钮提交执行了。

6. 确定聚类方法

在主对话框中，单击 Method 按钮，展开如图 13-18 所示的方法选择对话框。根据需要指定聚类方法、距离测度方法、对数值进行转换（标准化）的方法。参考本书附录 A。

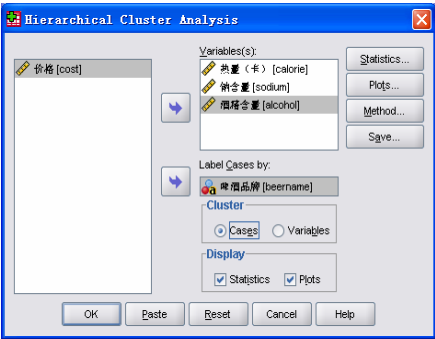


图 13-17 分层聚类分析主对话框

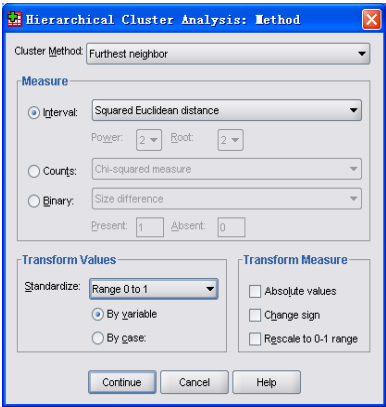


图 13-18 方法选择对话框

(1) 聚类方法选择。单击 Cluster Method 框中的向下箭头按钮，展开如图 13-19 所示的聚类方法菜单。

① Between-groups linkage，组间连接。合并两类的结果使所有的两两项对之间的平均距离最小。项对的两个成员分别属于不同的类。该方法中使用的是各对之间的距离，既非最大距离，也非最小距离。

② Within-groups linkage，组内连接。合并为一类后，类中的所有项之间的平均距离最小。该距离就是合并后的类中所有可能的观测量对之间的距离平方。

③ Nearest neighbor 项，最近邻法。首先合并最近的或最相似的两项，用两类间最近点间的距离代表两类间的距离。

④ Furthest neighbor，最远邻法。用两类间最远点的距离代表两类间的距离，也称为完全连接法。

⑤ Centroid clustering，重心聚类法。像计算所有各项均值之间的距离那样计算两类之间的距离，该距离随聚类的进行不断减小。

⑥ Median clustering，中位数法。以各类中的变量值中位数为类中心。

⑦ Ward's method 项，Ward 最小方差法。以方差最小为聚类原则。

(2) 对距离和相似性测度方法的选择。在 Measure 栏中指定的是用哪两点间的距离作为确定是否合并的距离。距离的具体计算方法还根据参与距离计算的变量类型，从以

下 3 种对话框选择其一，展开选择菜单后再进行具体方法的选择。这 3 个菜单分别对应于等间隔测度的变量（一般为连续变量）、计数变量（一般为离散变量）和二值变量。

① 对于等间隔测度的变量可在 **Interval** 的下拉菜单中选择连续变量距离测度的方法，如图 13-20 所示。这些方法是：

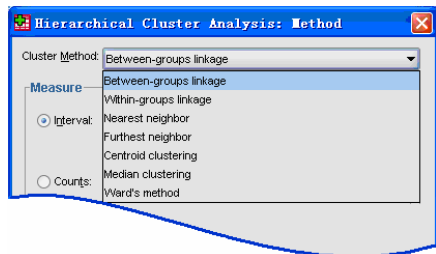


图 13-19 聚类方法下拉菜单

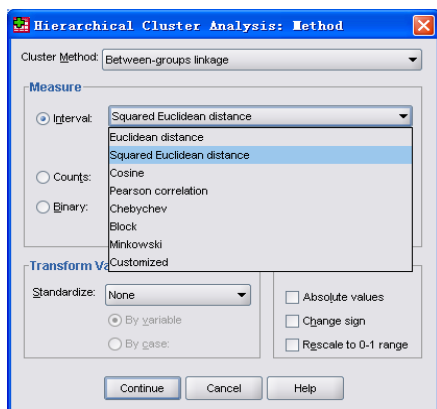


图 13-20 测度连续变量距离的方法

**Euclidean distance**（欧几里德距离，又称欧氏距离）、**Squared Euclidean distance**（欧氏距离平方）、**Cosine**（相似性测度）、**Pearson correlation**（皮尔逊相关）、**Chebychev**（切贝谢夫距离）、**Block**（布洛克距离）、**Minkowski**（明可斯基距离）。选择 **Minkowski** 后，还须输入乘方次数和开方次数  $p$ 。**Customized** 选项允许读者自定义距离计算方法，两项之间的距离用各项值之间差值的  $p$  次幂绝对值之和的  $r$  次方根表示，选择此项后，还须输入乘方次数  $p$  和开方次数  $r$ 。以上各选项的计算方法请见附录有关内容。

② 对于计数变量（离散变量）选择 **Count** 项。可在 **Count** 参数框下拉菜单中选择不相似性测度的方法：**Chi-square measure  $\chi^2$**  测度、**Phi-square measure  $\Phi^2$**  测度。各选项的计算方法请见附录有关内容。

③ 对于二值变量选择 **Binary** 项，在如图 13-21 所示的下拉菜单中选择距离或不相似性测度的方法。各方法的计算公式见附录有关内容。

首先应该明确，对二值变量，系统默认用 1 表示某特性的出现（或发生、存在等），用 0 表示某特性不出现（或不发生、不存在）。

在使用程序语句进行计算时如果指定两个参数，系统认为第一个参数表示某事件发生，第二个参数表示某事件不发生。如果只指定一个参数，系统认为该参数表示事件发生，其他值表示事件不发生。选项共

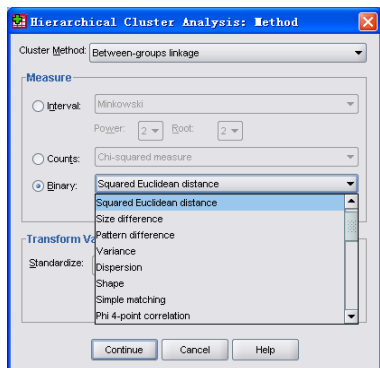


图 13-21 二值变量距离和相似性的测度方法

27 项。有关的约定和计算方法及解释请见附录 A 的有关内容。

从下拉表中的以上各项中选择一种测度方法。还可以改变表示某事件发生与不发生的值（或说某特性出现与不出现），在 **Present** 和 **Absent** 后面的矩形框中输入读者自己定义的值（当然应该与数据文件中有关的二元变量的值一致），定义后，系统将忽略其他值。如果不进行自定义，那么，1 代表某事件发生（**Present**），0 代表某事件不发生（**Absent**）。

(3) 选择标准化的方法。在 **Transform Values** 栏的 **Standardize** 标准化方法列表中选择，如图 13-22 所示。只有等间隔测度的数据（选择了 **Interval**）或计数数据（选择了 **Counts**）才可以进行标准化，对数据进行标准化的方法有：

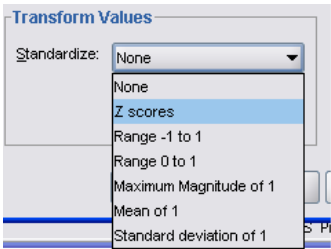


图 13-22 标准化方法菜单

- ① **None** 项，不进行标准化，是系统默认值。
- ② **Z scores** 项，把数值标准化到 Z 分数。
- ③ **Range -1 to 1** 项，把数值标准化到 -1 至 1 范围内。
- ④ **Range 0 to 1** 项，把数值标准化到 0 至 1 的范围内。
- ⑤ **Maximum magnitude of 1** 项，把数值标准化到最大值为 1。
- ⑥ **Mean of 1** 项，把数值标准化到均值为 1。
- ⑦ **Standard deviation of 1** 项，把数值标准化到单位标准差。

有关标准化方法的具体算法请看附录 A 的有关内容。

(4) 测度的转换方法选择，可选择的转换方法在 **Transform Measure** 栏中，有 3 种：

- ① **Absolute Values** 项，把距离值取绝对值。当数值符号表示相关方向，且只对负相关关系感兴趣时使用此方法进行变换。
- ② **Change sign**，把相似性值变为不相似性值或相反，用求反的方法使距离顺序颠倒。
- ③ **Rescale to 0~1 range**，通过先减去最小值然后除以范围的方法使距离标准化。

对已经按某种计算方法计算了相似性或不相似性测度的一般不再使用此方法进行转换。如果使用的是已经存在的矩阵，可以选择此类选项，对输入矩阵进行必要的转换。

第 6 步骤的四组选项选择完成后，按 **Continue** 按钮，返回到主对话框。

7. 选择要求输出的统计量

在主对话框中单击 **Statistics** 按钮，展开相应的 **Statistics** 对话框，如图 13-23 所示。指定要输出的统计量。

(1) **Agglomeration schedule**，要求做凝聚状态表。凝聚状态表显示聚类过程中每一步合并的两项（观测量与观测量、观测量与类、类与类）、被合并的两项之间的距离以及观测量或变量加入到一类的类水平，因此可以根据此表跟踪聚类的合并过程。由于最接近的两类先聚为一类，因此

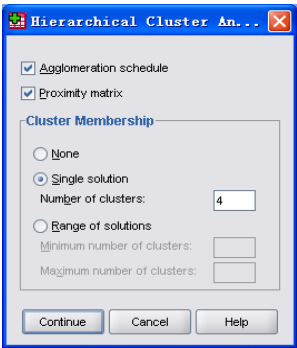


图 13-23 输出统计量对话框

可以通过聚类过程仔细地查看哪些观测量更接近一些。

(2) **Proximity matrix**, 要求输出各项间的距离矩阵。以矩阵形式给出各项之间的距离或相似性测度值。产生什么类型的矩阵(相似性矩阵或不相似性矩阵)取决于在 **Method** 对话框中 **Measure** 栏中的选择。注意: 如果项数很大(观测量数或变量数很大)该选项产生的输出量也会很大。

(3) **Cluster Membership**, 类成员栏。要求显示每个观测量被分派到的类(即分类结果, 各观测量属于哪一类)或显示若干步凝聚过程。可以用下面的选项进一步选择:

① **None**, 不显示类成员表, 是系统默认的。

② **Single solution**, 要求聚为指定类数时, 列出各观测量所属的类。在 **Number of cluster** 选项右侧的矩形框中输入限定显示的类数, 该数值必须是大于 1, 小于等于参与聚类的观测量或变量总数的整数, 例如在矩形框中输入数字“3”, 则会在输出窗口中显示聚为三类时每个观测量属于三类中的哪一类。

③ **Range of solutions**, 要求列出某个范围中每一步聚类过程和各观测量所属的类。

- 在 **Minimun number of clusters** 后面输入一个最小类数值;
- 在 **Maximun number of clusters** 后面输入最大类数值。

这两个数值必须是不等于 1 的正整数, 最大类数值不能大于参与聚类的观测量数或变量总数, 例如, 指定此选项并且在左右两个矩形框中分别输入了“3”和“5”这两个数值。将在输出窗口中显示三个结果: 观测量(或变量)被聚为 3 类、4 类、5 类时各观测量(或变量)被分派到哪一类。

以上内容所涉及的变量或观测量取决于在主对话框中的 **Cluster** 栏中选择的是 **Cases** (观测量) 还是 **Variables** (变量)。

## 8. 选择统计图表

在分层聚类分析的主对话框中, 单击 **Plot** 按钮, 展开如图 13-24 所示统计图表输出子对话框。

(1) **Dendrogram** 要求输出树形图。

(2) **Icicle**, 输出冰柱图栏, 对于生成什么样的冰柱图还可以进一步用以下选项确定:

① **All clusters**, 聚类的每一步都表现在图中。可用此种图查看聚类的全过程, 但如果参与聚类的个体很多会造成图过大, 没有必要。

② **Specified range of clusters**, 指定显示的聚类范围。当选择此项时, 在 **Start cluster** 框中输入要求显示聚类过程的起始步数, 在 **Stop cluster** 框中输入中止于哪一步, 在 **By** 框中输入两步之间的增量, 输入到矩形框中的数字必须是正整数。例如, 输入的结果是: **Startcluster: 3**,

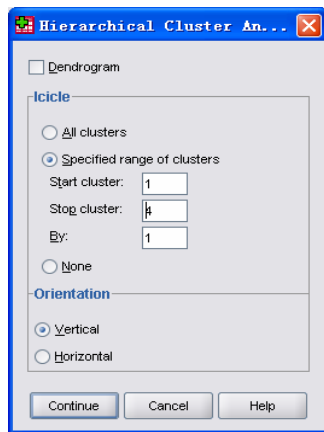


图 13-24 统计图对话框

Stopcluster: 10 By: 2。生成的冰柱图从第三步开始，显示第三、五、七、九步聚类的情况。

③ None，不生成冰柱图。

(3) 对于显示方向可以在 Orientation 栏中确定。

① Vertical，显示纵向的冰柱图。

② Horizontal，显示水平的冰柱图。

### 9. 生成新变量的选项

聚类分析的结果可以用新变量保存在工作数据文件中。单击主对话框的 Save 按钮，展开如图 13-25 所示的对话框。可以看出只能生成一个表明参与聚类的个体最终被分配到哪一类的新变量。通过对话框可以选择是否建立新变量和建立的新变量含义。

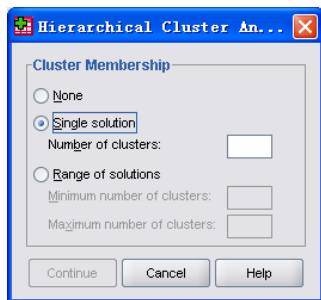


图 13-25 新变量选择对话框

(1) None，不建立新变量。

(2) Single solution，即单一结果，生成一个新变量，表明每个个体聚类最后所属的类。在 Number of clusters:后面的矩形框中指定类数，如果输入 5，则新变量值的范围为 1~5。

(3) Range of solutions，即指定范围内的结果。生成若干个新变量，表明聚为若干个类时，每个个体聚类后所属的类。把表示从第几类显示到第几类的数字分别输入到 Minimum number of clusters 后和 Maximum number of clusters 后面的矩形框中，例如输入结果是“Minimum

number of clusters:4, Maximum number of clusters: 6”，在聚类结束后在数据窗口中原变量后面增加了 3 个新变量分别表明分为 4 类时、分为 5 类和分为 6 类时的聚类结果。即各观测量分别属于哪一类。

## 13.4.3 样本分层聚类分析实例

【例 4】例 1 数据文件 data13-03 为一组有关 12 盎司啤酒成分和价格的数据，变量包括：beername（啤酒品牌）、calorie（热量卡路里）、sodium（钠含量）、alcohol（酒精含量）、cost（价格）。要求根据 12 盎司啤酒的各成分含量及 12 盎司啤酒价格对 20 种啤酒进行分类。

(1) 读取数据文件后，操作步骤如下：

① 按 Analyze→Classify→Hierarchical Cluster 顺序，单击菜单项，展开 Hierarchical Cluster Analysis 分层聚类分析对话框。

② 选择 calorie、sodium、alcohol、cost 这 4 个变量为分析变量，移到 Variable(s)框中。选择 beername 作为标识变量，移到 Label Cases 框中。

③ 选择 Q 型聚类。在 Cluster 栏中选择系统默认的 Cases 项。

④ 选择要求输出的统计量。在 Display 栏中选中 Statistics 复选项，单击 Statistics

按钮，展开统计量选择对话框，进行下列选择：

选中 **Agglomeration schedule**，要求输出凝聚状态表。

选中 **Proximity matrix**，要求输出距离矩阵。

在 **Cluster membership** 栏中选择 **Single solution**，并在其后的小矩形框中输入“4”，即要求聚类进行到把所有观测量分为 4 类时，显示每个观测量所属的类。

⑤ 选择聚类方法，单击主对话框中的 **Method** 按钮，打开 **Method** 对话框。

- Cluster Method 参数框中选择聚类法，本次运算选择 **Fuethest neighbor** 最远邻法项。

- Measure 栏中选择 **Interval**，下拉列表中选择 **Squared Euclidean distance** 项。对等间隔测度的变量使用欧氏距离平方作为类间距离。

- Transform Value 栏中选择标准化方法。在 **Standardize** 列表中选择 **range 0 to 1** 项。

选择 **By variable** 项。注意：必须指定标准化方法。因为 4 个分析变量单位不同。

⑥ 选择要求显示的统计图。在主对话框中的 **Display** 栏中选中 **Plots** 复选项，单击 **Plots** 按钮，展开统计图表选择对话框，进行下列选择：

- 选中 **Dendrogram**，要求做树形图。

- 要求作冰柱图，但不要求把聚类全过程都表现在图上，而是只表现聚为 4 类的过程。因此在 **Icicle** 栏中选择 **Specified range of clusters** 项，并在下面的三个小矩形框中填入数字：**Start cluster: 1**，**Stop cluster: 4**，**By: 1**。在 **Orientation** 内选择 **Vertical** 项，即纵向作图。

⑦ 选择要存入数据文件的新变量。在主对话框中按 **Save** 按钮，选择 **Cluster Membership** 框中的 **Single solution** 项，在 **Number of clusters** 后的小矩形框中输入 4，即要求在工作数据文件中建立新变量，当把所有观测量分为 4 类时，该变量值表明每个观测量被分派到的类号。

⑧ 在主对话框中按 **OK** 按钮，直接提交运行。

(2) 程序的命令语句

```
PROXIMITIES  calorie sodium alcohol cost                                ①
/MATRIX OUT  ('C:\WINDOWS\TEMP\ spss1104\spssclus.tmp')  /VIEW= CASE

/MEASURE= SEUCLID  /PRINT  NONE
/ID= beername  /STANDARDIZE= VARIABLE RESCALE .

CLUSTER                                              ②
/MATRIX IN  ('C:\WINDOWS\TEMP\ spss1104\spssclus.tmp')
/METHOD COMPLETE  /ID=beername  /PRINT SCHEDULE CLUSTER(4)
/PRINT DISTANCE  /PLOT DENDROGRAM VICICLE(1,4,1)  /SAVE CLUSTER (4) .
ERASE FILE= 'C:\WINDOWS\TEMP\ spss4294656823\spssclus.tmp'.
```

### (3) 程序解释如下:

进行分层聚类分析由两个 SPSS 过程完成, PROXIMITIES 过程对原始数据进行变换, 即标准化, 并对距离进行计算。然后由 Cluster 过程进行观测量聚类。

① PROXIMITIES 过程首先列出分析变量: alcohol、calorie、cost、sodium。

/MATRIX OUT 子命令指定了计算结果存入 C:\WINDOWS\TEMP 中由系统指定并自动建立的临时代号子文件夹中, 临时文件名为 spssclus.tmp。这是一个距离矩阵。

/VIEW=CASE 子命令, 指定近似计算的目的是要对观测量进行聚类。

/MEASURE=SEUCLID 子命令, 指定距离测度使用欧氏距离平方。

/ID=beername 子命令指定用变量 beername 即啤酒名作为标识变量。

/STANDARDIZE=VARIABLE RESCALE 子命令指定将参与分析的变量值标准化到 0 至 1 范围。RESCALE 是 Range 0 to 1 方法的关键字。

/PRINT NONE 子命令指定计算的不相似性矩阵不输出到 Viewer 窗口。

### ② CLUSTER 过程使用的子命令

/MATRIX IN 子命令把由 PROXIMITIES 输出的矩阵作为进行聚类分析的依据, 从存储文件中读入。

/METHOD COMPLETE 子命令指定观测量聚类采用最远邻 (即完全连接) 方法。该子命令对应于 Method 对话框中指定的 Furthest Neighbor 选项。

/ID=beername 子命令指定变量名 beername 作为标识变量。

/PRINT SCHEDULE CLUSTER(4) 和 /PRINT DISTANCE 两个 PRINT 子命令要求输出凝聚表 and 把观测量聚为 4 类的聚类结果表。可供研究聚类过程和确定最终聚为几类比较合理。运行 CLUSTER 后面括号中取不同数值的程序, 可以比较聚为不同类的结果。

/PLOT DENDROGRAM VICICLE(1,4,1) 子命令要求输出两种统计图, DENDROGRAM 要求输出反映聚类全过程的树形图; VICICLE(1,4,1) 要求输出把观测量聚为 4 类的纵向冰柱图, 要求从聚为 1 类到聚为 4 类逐步作图。

/SAVE CLUSTER(4) 子命令要求在工作数据文件中建立新变量, 其数值为各观测量在聚为 4 类时所属的类号。

③ ERASE FILE 命令要求把存于 C:\WINDOWS\TEMP 目录下的中间结果临时文件 spssclus.tmp 删除。

(4) 在输出窗口中输出结果见表 13-16 至表 13-18、图 13-26、图 13-27。

### (5) 输出结果解释

表 13-16 是欧氏不相似性系数平方矩阵, 它是  $20 \times 20$  方阵。行顶、最左列均是啤酒名, 在行列交叉点上这是两种啤酒 4 个变量的欧氏距离的平方和, 体现的是不相似性, 数值越大, 两种啤酒越不相似。如果读者使用该数据集数据进行同样的分析, 表中内容会稍有不同, 这是因为本书作者对 Viewer 窗口中的表格进行了编辑, 列宽的调整有时会显示的有效位数。

表 13-17 是 Cluster 过程的输出。由于在 Statistics 选项中选择 Agglomeration schedule, 输出在 Output 窗口中为一个表明聚类过程的表, 其中:

- Stage, 聚类步顺序号。(Clusters Combined) Cluster1, Cluster2, 是该步被合并的两类中的观测量号。

表 13-16 欧氏不相似性系数平方矩阵

Case	Proximity Matrix																			
	Squared Euclidean Distance																			
	1: Bud weis er	2: Schli tz	3: Ione nbra u	4: Kron enso urc	5: Hein eken	6: Old- milna ukee	7: Aucs berg er	8: Strch s-bo hemi	9: Mille r-lite	10: Sude iser-l ich	11: Coors	12: Coors licht	13: Mich elos- lich	14: Secrs	15: Kkirin	16: Pabst -extra -l	17: Ham ms	18: Heile man s-old	19: Olym pia-g old-	20: Schli te-lig ht
1: Budweiser	.000	.111	.062	.724	.570	.140	.198	.147	.358	.556	.023	.213	.193	.391	.855	1.069	.014	.061	1.109	.530
2: Schlitz	.111	.000	.090	.665	.623	.249	.098	.230	.745	.886	.161	.591	.376	.467	.926	1.714	.183	.164	1.708	.933
3: Ionenbrau	.062	.090	.000	.390	.339	.337	.267	.348	.364	.482	.039	.301	.123	.323	.532	1.332	.104	.206	1.142	.475
4: Kronensourc	.724	.665	.390	.000	.071	1.451	1.054	1.308	.815	.776	.589	.885	.418	.385	.054	2.269	.800	1.037	1.531	.756
5: Heineken	.570	.623	.339	.071	.000	1.272	.936	1.026	.682	.729	.471	.653	.345	.155	.059	1.899	.612	.801	1.331	.656
6: Old-milnaukee	.140	.249	.337	1.451	1.272	.000	.222	.130	.661	.930	.228	.457	.555	.929	1.672	1.162	.149	.114	1.497	.934
7: Aucsberger	.198	.098	.267	1.054	.936	.222	.000	.137	1.04	1.358	.326	.805	.709	.630	1.354	2.086	.297	.114	2.239	1.314
8: Strchs-bohem	.147	.230	.348	1.308	1.026	.130	.137	.000	.867	1.201	.283	.540	.643	.557	1.496	1.416	.168	.027	1.786	1.152
9: Miller-lite	.358	.745	.364	.815	.682	.661	1.041	.867	.000	.087	.222	.065	.122	.791	.741	.540	.292	.638	.288	.027
10: Sudeiser-lich	.556	.886	.482	.776	.729	.930	1.358	1.201	.087	.000	.363	.210	.132	.953	.703	.556	.473	.951	.196	.050
11: Coors	.023	.161	.039	.589	.471	.228	.326	.283	.222	.363	.000	.141	.087	.394	.685	.948	.026	.156	.873	.347
12: Coorslicht	.213	.591	.301	.885	.653	.457	.805	.540	.065	.210	.141	.000	.128	.572	.823	.443	.139	.388	.395	.148
13: Michelos-lich	.193	.376	.123	.418	.345	.555	.709	.643	.122	.132	.087	.128	.000	.428	.434	.810	.167	.455	.538	.153
14: Secrs	.391	.467	.323	.385	.155	.929	.630	.557	.791	.953	.394	.572	.428	.000	.395	1.695	.412	.451	1.496	.870
15: Kkirin	.855	.926	.532	.054	.059	1.672	1.354	1.496	.741	.703	.685	.823	.434	.395	.000	2.068	.893	1.199	1.283	.641
16: Pabst-extra-l	.069	1.714	1.33	2.269	1.899	1.162	2.086	1.416	.540	.556	.948	.443	.810	1.695	2.068	.000	.847	1.314	.256	.607
17: Hamms	.014	.183	.104	.800	.612	.149	.297	.168	.292	.473	.026	.139	.167	.412	.893	.847	.000	.086	.927	.455
18: Heilemans-old	.061	.164	.206	1.037	.801	.114	.114	.027	.638	.951	.156	.388	.455	.451	1.199	1.314	.086	.000	1.535	.882
19: Olympia-gold	.109	1.708	1.14	1.531	1.331	1.497	2.239	1.786	.288	.196	.873	.395	.538	1.496	1.283	.256	.927	1.535	.000	.217
20: Schlite-light	.530	.933	.475	.756	.656	.934	1.314	1.152	.027	.050	.347	.148	.153	.870	.641	.607	.455	.882	.217	.000

This is a dissimilarity matrix

- Coefficient, 距离测度值。表明不相似性的系数。由于选择了欧氏距离平方作为距离测度, 因此从表中可以看出数值较小的两项 (两个观测量、两类或观测量与一类) 比数值较大的两项先合并。第一步是第 1 个观测量与第 17 个观测量合并; 第二步为第 1 个和第 11 个观测量合并。这样两步合并了三个观测量到一类。

- Stage Cluster First Appears, 合并的两项第一次出现的聚类步序号。Cluster1 和 Cluster2 值均为 0 的是两个观测量合并, 其中有一个为 0 的是观测量与类合并; 两个值均为非 0 值的是两个类合并, 如第 6 步为第 4 观测量与第 5 观测量合并, 而第 5 观测量在第 5 步已经与第 15 观测量合并为一类了, 因此, 此项值的 5 表示观测量 4 与第 5 步形成的类归并为一类。

- Next stage, 此步合并结果在下一步合并时的步序号。

选择不同的标准化算法, 距离矩阵会有所不同。选择不同的算法和对距离的测度方法, 聚类过程是不同的, 因而聚类结果也会有所区别。

表 13-18 聚类结果表明各观测量分别分到四类中的哪一类。

图 13-26 所示冰柱图。从缺少 “×” 处分界, 可以看出四类的划分。如果作出全观



测量图，可以清楚地看到所有观测量最后聚为一类的全过程。由于选项 VICICLE(1,4,1) 指定的参数，图中显示了从聚为一类到聚四类的全过程。观其形，如果是反映聚类全过程的图形，无论是作为观测量标识的自上至下列出的啤酒名，还是反映聚类过程的图形本身都酷似自上而下的冰柱，因而由此命名。从此图中可以清楚地看到哪几种啤酒被归为一类，从而得出最后的分类结论。

表 13-17 聚类的凝聚过程表

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	17	.014	0	0	2
2	1	11	.024	1	0	8
3	8	18	.027	0	0	10
4	9	20	.027	0	0	7
5	4	15	.054	0	0	6
6	4	5	.065	5	0	16
7	9	10	.069	4	0	12
8	1	3	.069	2	0	14
9	2	7	.098	0	0	13
10	6	8	.122	0	3	13
11	12	13	.128	0	0	12
12	9	12	.138	7	11	17
13	2	6	.186	9	10	14
14	1	2	.197	8	13	18
15	16	19	.256	0	0	17
16	4	14	.312	6	0	18
17	9	16	.459	12	15	19
18	1	4	.801	14	16	19
19	1	9	.860	18	17	0

表 13-18 共分为四类的聚类结果

Cluster Membership	
Case	4 Clusters
1: Budweiser	1
2: Schlitz	1
3: Ionenbrau	1
4: Kronensourc	2
5: Heineken	2
6: Old-milnaukee	1
7: Aucsberger	1
8: Strchs-bohemi	1
9: Miller-lite	3
10: Sudeiser-lich	3
11: Coors	1
12: Coorslicht	3
13: Michelos-lich	3
14: Secrs	2
15: Kkirin	2
16: Pabst-extra-l	4
17: Hamms	1
18: Heilemans-old	1
19: Olympia-gold-	4
20: Schilte-light	3

Vertical Icicle																				
Number of cluster	Case																			
	19: Olympia-gold-	16: Pabst-extra-l	13: Michelos-lich	12: Coorslicht	10: Sudeiser-lich	20: Schilte-light	9: Miller-lite	14: Secrs	5: Heineken	15: Kkirin	4: Kronensourc	18: Heilemans-old	8: Strchs-bohemi	6: Old-milnaukee	7: Aucsberger	2: Schlitz	3: Ionenbrau	11: Coors	17: Hamms	1: Budweiser
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

图 13-26 聚为四类的冰柱图

图 13-27 是反映聚类全过程的树形图。可以在此图上用一把尺子垂直方向放在图上左右移动，与尺子相交的每一根横线就是一类。每根横线左端与之联系的各观测量就是分到该类的成员。大致观察一下，决定如何分类合适。图上方的数字是按距离比例进行重新标定的结果，不影响对分类结果的观察与结论。可以看出分为 2 类、3 类或 4 类时，类间距离比较大，说明各类的特点比较突出，对各类啤酒容易定义。分为 5 类以上，有些类间的区别不很明显。

图 13-28 是工作数据文件的一部分，其中最右面的一列是新变量 clu4\_1，其值表明聚为四类时各观测量所属类的类号。反复运行定义语句：“/SAVE CLUSTER(n)”，n 改变为不同值时的程序，得到形式为 clun\_1 的变量，比较 n 为各种值时的分类结果，可以确

定具体的分类结果。“-”后面的数字表示是在当前的 SPSS 期间第几次运行该过程生成的新变量。

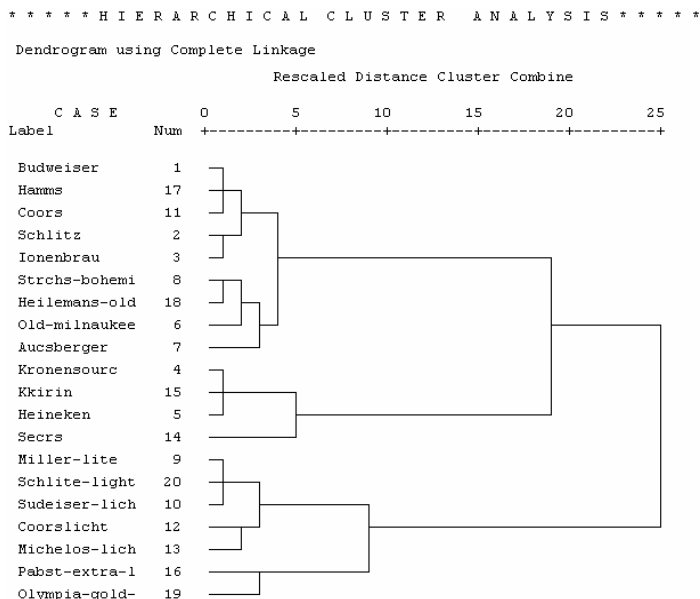


图 13-27 聚类树形图（重新标定到 0~25）

比较表 13-18 与图 13-28，可以看出，在工作数据文件中建立的新变量也表明各观测量所属的类，因此可以根据需要选择一个即可。在 Save 对话框中指定了建立新变量，可以不必在 Statistics 对话框中指定 Cluster Membership 的选项。在程序中的子命令/PRINT CLUSTER(4)与子命令/SAVE CLUSTER(4)中的两个选项 CLUSTER(4)只要一个即可，结果只是查看分类结果的位置不同而已。

比较时应该注意，变量 clun\_1 的值只是序号。采用相同的 n 值，不同的聚类方法和不同的测度不相似性（距离）的算法，结果可能有区别，例如对于“/SAVE CLUSTER(4)”第一次、第二次、第三次执行采用不同聚类方法和不同测度距离的方法的程序，会建立 clu4\_1, clu4\_2, clu4\_3, …，的变量，便于比较、得出结论。

【例 5】例 2 使用另一些选项的程序与输出。

应该说明的是分类是根据特定的目的进行的。对于同样一些观测量，不同的分类目

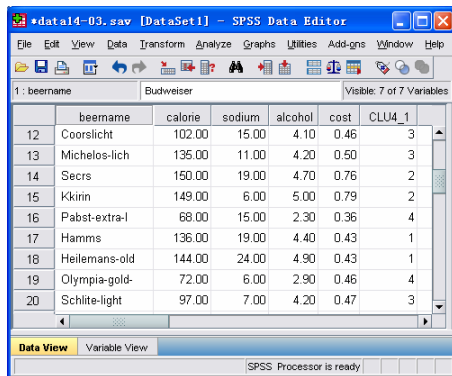


图 13-28 加入了新变量的工作数据文件

的, 使用反映不同特征的变量, 分类的结果就不相同。同一分类目的, 根据不同的实际需要, 也可以分成不同的类数。因此可以在使用 CLUSTER 过程时指定不同的参数, 对不同的结果进行比较。以便得出符合实际需要的结论。

### (1) 操作步骤

①~④步操作与例 1①~④步操作相同。

⑤ 选择显示的统计量。在 Statistics 对话框中 Agglomeration schedule 项和 Proximity Matrix 项的选项与例 1 相同。

在 Cluster Membership 栏中选择 Range of solutions 项, 并在其后输入 3 和 5, 即要求聚类进行到把所有观测量分为 3 类、4 类、5 类时, 显示每一个观测量所属的类。

⑥在 Method 对话框中, 选择聚类方法。Cluster Method 框中, 选择 Centroid clustering 在 Standardize 框中, 选择 Zscores。

⑦ 选择要求显示的统计图。

Dendrogram 选项, 要求作树形图。

要求作冰柱图, 但不要求把聚类全过程都表现在图上, 而是只表现聚为 4 类的过程, 因此在 Icicle 栏中选择 Specified range of clusters 项, 并在下面的三个小矩形框中填入数字, 即 Start cluster 框中输入 3, Stop cluster 框中输入 10, By 框中输入 2, 即要求作冰柱图表明聚为 3、5、7、9 类时的分类情况。在 Orientation 内选择作图方向, 仍选择 Vertical 项, 即纵向图。

⑧ 选择要存入数据文件的新变量。

在主对话框中按 Save 按钮, 展开相应的对话框。选择 Cluster Membership 框中的 Range of solution 项。在 Minimum number of cluster 后的小矩形框中输入 2, Maximum number of cluster 后的小矩形框中输入 6, 即要求在工作数据文件中建立 5 个新变量。当把所有观测量分为 2 类、3 类、4 类、5 类和 6 类时对应变量值表明每个观测量被分派到的类号。

(2) 形成由命令语句组成的 SPSS 程序如下:

```
PROXIMITIES  calorie sodium alcohol cost
```

```
  /MATRIX OUT ('E:\DOCUME~1\Wendylu\LOCALS~1\Temp\spss856\spssclus.tmp')
```

```
  /VIEW= CASE  /MEASURE= SEUCLID  /PRINT  NONE  /ID= beername
```

```
  /STANDARDIZE= VARIABLE Z .
```

```
CLUSTER
```

```
  /MATRIX IN  ('E:\DOCUME~1\Wendylu\LOCALS~1\Temp\spss856\spssclus.tmp')
```

```
  /METHOD CENTROID  /ID=beername  /PRINT SCHEDULE CLUSTER(3,5)
```

```
  /PLOT DENDROGRAM VICICLE(3,10,2)  /SAVE CLUSTER(2,6) .
```

```
ERASE FILE= 'E:\DOCUME~1\Wendylu\LOCALS~1\Temp\spss856\spssclus.tmp'.
```

与前一个程序不同之处是:

PROXIMITIES 中的子命令语句, /STANDARDIZE= VARIABLE Z 要求对变量进行 Z 分数法的标准化。

CLUSTER 过程中的子命令语句/PRINT SCHEDULE CLUSTER(3,5)要求输出形成 3、4、5 类的过程; /PLOT DENDROGRAM VICICLE(3,10,2)要求作树形图, 冰柱图只显示聚成 3、5、7、9 类的结果; /SAVE CLUSTER(2,6) 要求保存 2、3、4、5、6 类的结果, 各观测量被分到哪一类的新变量。

(3) 结果输出见表 13-19、图 13-29、图 13-30。

① 由于选择的计算不相似性系数(距离)的方法与前例相同, 但标准化方法不同因此不相似性矩阵输出结果与表 13-19 也不同。聚类方法选择与前例相同, 但由于标准化方法不同聚类的凝聚过程与表 13-20 也不同。为节省篇幅, 读者自己观察这两项输出。

表 13-19 不同标准化方法聚为 3、4、5 类的结果

Cluster Membership (Z分数)				Cluster Membership (标准化0-1)			
Case	5 Clusters	4 Clusters	3 Clusters	Case	5 Clusters	4 Clusters	3 Clusters
1: Budweiser	1	1	1	1: Budweiser	1	1	1
2: Schlitz	1	1	1	2: Schlitz	1	1	1
3: Ionenbrau	1	1	1	3: Ionenbrau	1	1	1
4: Kronensourc	2	2	2	4: Kronensourc	2	2	2
5: Heineken	2	2	2	5: Heineken	2	2	2
6: Old-milnaukee	1	1	1	6: Old-milnaukee	1	1	1
7: Auscberger	1	1	1	7: Auscberger	1	1	1
8: Strchs-bohemi	1	1	1	8: Strchs-bohemi	1	1	1
9: Miller-lite	3	3	1	9: Miller-lite	3	3	3
10: Sudeiser-lich	3	3	1	10: Sudeiser-lich	3	3	3
11: Coors	1	1	1	11: Coors	1	1	1
12: Coorslicht	3	3	1	12: Coorslicht	3	3	3
13: Michelos-lich	3	3	1	13: Michelos-lich	3	3	3
14: Secrs	4	2	2	14: Secrs	4	2	2
15: Kkirin	2	2	2	15: Kkirin	2	2	2
16: Pabst-extra-l	5	4	3	16: Pabst-extra-l	5	4	3
17: Hamms	1	1	1	17: Hamms	1	1	1
18: Heilemans-old	1	1	1	18: Heilemans-old	1	1	1
19: Olympia-gold-	5	4	3	19: Olympia-gold-	5	4	3
20: Schlite-light	3	3	1	20: Schlite-light	3	3	3

		Vertical Icicle																			
		Case																			
Number of clusters		19: Olympia-gold-	16: Pabst-extra-l	14: Secrs	15: Kkirin	5: Heineken	4: Kronensourc	13: Michelos-lich	12: Coorslicht	10: Sudeiser-lich	20: Schlite-light	9: Miller-lite	7: Auscberger	2: Schlitz	18: Heilemans-old	8: Strchs-bohemi	6: Old-milnaukee	3: Ionenbrau	11: Coors	17: Hamms	1: Budweiser
3		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
5		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
7		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
9		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

图 13-29 第 3、5、7、9 步聚类的纵向冰柱图

聚类结果见表 13-22。左面的表与右面的表是使用不同标准化的方法产生的不同结果。都是聚为 3、4、5 类的结果。聚类结果是有差别的。

② 图 13-29 为使用 Z 分数法对原始变量进行标准化, 反映聚类过程的冰柱图, 从冰柱图可以很清晰地看出如果分三类, 用观测量序号表示则:

第一类包括的是编号为 1、2、3、6、7、8、11、17、18 的啤酒。  
第二类包括的是编号为 4、5、9、10、12、13、14、15、20 的啤酒。  
第三类包括的是编号为 19、16 的啤酒。

如果分为五类，是第二类分成两类：4、5、15、14 和 9、20、10、12、13 各聚成单独的一类。第一类分为 1、17、11、2、3 和 6、8、18、7 两类。

以此类推读者自己可以看出分为七类或分为九类的各啤酒的分类结果。对照不相似性系数矩阵，会对聚类的原理有更进一步的理解。

③ 图 13-30 所示在工作数据文件中建立的新变量共有 5 个。

clu6\_1、clu5\_1、clu4\_1、clu3\_1、clu2\_1，分别表示当前的 SPSS 期间第一次运行分层聚类过程，如果分为 6 类或 5 类、4 类、3 类、2 类时各观测量所属类别。每个变量的值为相应的观测量分到类的代码。这里的第几类没有任何含义，只是标记。至于哪类是什么特征还需认真分析工作数据窗口中的原始数据、新生成的分类变量和专业知识来确定，并可以进一步对每一类进行命名。

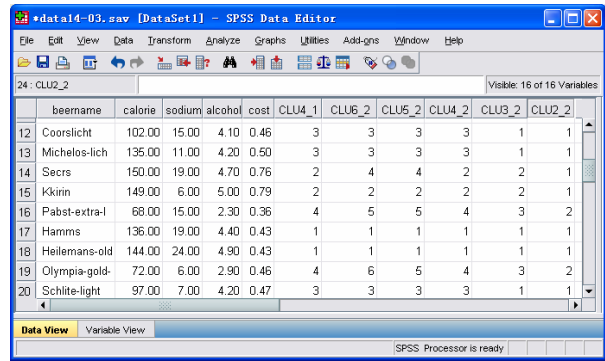


图 13-30 工作数据文件增加五个结果变量

观测量的分类特征。应该在变量选择上下工夫。

13.4.4 变量聚类概述

1. 变量聚类的概念

变量聚类也称 R 型聚类，是一种降维的方法，用于在变量众多时寻找有代表性的变量，以便在用少量、有代表性的变量代替大变量集时，损失信息很少。这种方法在人类学、动物学、医学和工业生产中以及市场分析中都得到应用，例如人种分类、动植物分类等，往往要测量许多表明形态特性的变量值。某些变量之间有很强的相关性，找出一个变量可以代替一系列与其相关的变量，则可大大减少工作量，节省测量时间，但不会影响分类的结果。因此，在分类学中选择变量是一步很重要的工作。变量聚类是选择变

注意：SPSS 提供了众多的聚类方法和标准化方法。分析数据时，都是人为选定某种方法。不同的聚类算法和不同的对变量进行标准化的方法都会对聚类结果有影响。而类本身就是针对某一研究目的而进行的。因此在作结论时，一定要结合专业知识、研究目的，同时认真观察原始数据特征，审慎地得出结论。并对分成的各类进行命名。如果不同方法得出的结果差别很大，说明聚类变量选择的不是真正反映

量的很实用的方法之一。另外进行回归分析时也需要首先降维以便找出相互独立的变量。

## 2. 选择代表指标的方法

聚类结束后, 各类变量中选择哪个变量作为代表变量呢? 典型指标的选择主要根据专业知识, 同时根据下列原则综合确定代表变量。考察在一类指标中:

(1) 最有代表性的变量。

(2) 最容易测得的变量。例如, 测试仪器容易得到、仪器便宜、测试对象容易接受、指标数据容易测得准确等各方面因素。例如医学研究中, 尿量虽然容易测得, 但 24 小时尿量不易收全, 就不易准确。此时就应该考虑, 在与之聚为一类的变量中, 其他变量中是否有更好的代替者。

(3) 如果从专业角度不好确定, 还可以通过进行进一步计算来确定。

例如,  $x_1$ 、 $x_2$ 、 $x_3$ 、 $x_4$  这 4 个指标已经根据 R 型聚类结果聚为一类。

① 计算每个指标的相关指数, 公式为

$$\bar{R}_j^2 = \frac{\sum r^2}{m_j - 1}$$

其中:  $r$  为指标  $x_j$  与同类中其他指标间的相关系数;  $m_j$  为指标  $x_j$  所在类的指标个数。

② 对  $x_1$ 、 $x_2$ 、 $x_3$ 、 $x_4$  这 4 个指标计算  $\bar{R}_1^2$ 、 $\bar{R}_2^2$ 、 $\bar{R}_3^2$ 、 $\bar{R}_4^2$ , 比较这 4 个值, 最大一个相关指数对应的变量, 可以选做典型指标。

3. SPSS 使用 CLUSTER 过程对变量进行聚类。操作步骤与方法均与使用 CLUSTER 过程对观测量进行聚类是相同的。不同点在于:

(1) 在主对话框 Hierarchical Cluster Analysis 中的 Cluster 框中选择 Variable 项。

(2) 主对话框中的 Save 按钮为灰色, 不能单击。因为变量聚类不建立新变量。

## 13.4.5 变量聚类分析实例

【例 6】啤酒分类的问题中, 是否有必要使用 4 个变量进行分析? 可以用变量聚类方法解决这个问题。数据仍为 data13-03。

### (1) 操作步骤

① 按 Analyze→Classify→Hierarchical Cluster 顺序单击菜单项, 打开主对话框。

② 选择 4 个变量: colore (热量)、sodium (钠含量)、alcohol (酒精含量)、cost (价格) 为分析变量, 移到 Variables 栏中。在 Cluster 栏中选择 Variable 项。

③ 单击 Method 按钮打开相应对话框。

- 在 Cluster Method 栏中选择 Furthest neighbor 作为聚类方法。

- 在 Measure 栏中选择 Interval 中的 Pearson Correlation 皮尔逊相关作为测度变量间相似性的方法。也因此在 Transform Values 栏选择 Standardize 中的 None 不进行标准化。

④ 单击 Plot 按钮, 打开相应的对话框, 选择 Dendrogram 项。在 Icicle 栏中选择 All Clusters, 要求在冰柱图中反映聚类全过程。

⑤ 单击 Statistics 按钮，打开相应的对话框。选择 Proximity Matrix，要求显示相关系数矩阵。

(2) 在主对话框中单击 Paste 按钮，得到下面的程序语句：

```
PROXIMITIES  calorie sodium alcohol cost

/MATRIX OUT  ('C:\WINDOWS\TEMP\spss3016\spssclus.tmp')  /VIEW= VARIABLE
/MEASURE= CORRELATION /PRINT  NONE  /STANDARDIZE=VARIABLE NONE.

CLUSTER

/MATRIX IN  ('C:\WINDOWS\TEMP\spss3016\spssclus.tmp') /METHOD COMPLETE
/PRINT DISTANCE  /PLOT DENDROGRAM VICICLE.

ERASE FILE= 'C:\WINDOWS\TEMP\spss3016\spssclus.tmp'.
```

(3) 程序解释

在 PROXIMITIES 命令中给出参与分析的变量表。MATRAX 子命令要求计算的相关矩阵。并保存为临时文件，C:\WINDOWS\TEMP\ spss4294658765\spssclus.tmp 指定了文件保存路径和临时文件名。该矩阵文件供 Cluster 过程分析时使用。VIEW 子命令要求对变量进行聚类。STANDARDIZE=NONE 命令不要求对变量进行标准化。

CLUSTER 命令中的 MATRIX IN 子命令要求分析 PROXIMITIES 生成的包含相关矩阵的数据文件。METHOD 子命令要求使用完全连接法即最远邻法进行聚类。PRINT 子命令要求输出聚类过程表和表明变量距离的相关矩阵。PLOT 子命令要求作树形图和全过程的纵向冰柱图。ERASE FILE 命令删除临时数据文件。

(4) 运行结果见表13-20和图13-31、图13-32。省略综合信息表和聚类过程表。

表13-20 变量的相关矩阵

Proximity Matrix				
Case	Matrix File Input			
	热量 (卡)	钠含量	酒精含量	价格
热量 (卡)	1.000	.429	.903	.291
钠含量	.429	1.000	.337	-.444
酒精含量	.903	.337	1.000	.345
价格	.291	-.444	.345	1.000

Vertical Icicle						
Number of clusters	Case					
	价格	钠含量	酒精含量	热量 (卡)		
1	×	×	×	×	×	×
2	×	×	×	×	×	×
3	×	×	×	×	×	×

图 13-31 冰柱图

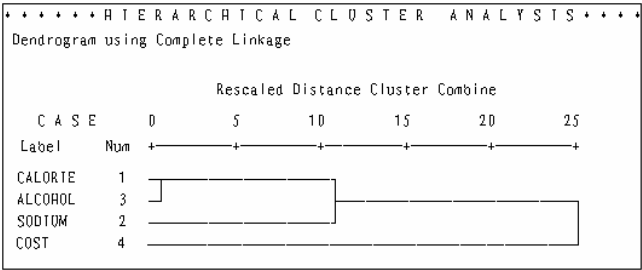


图 13-32 变量聚类的树形图

无论从相关矩阵还是冰柱图、树形图都可以看出热量和酒精含量两个变量相关系数最大, 首先聚为一类。从整体看, 聚为三类是比较好的结果。至于热量和酒精含量选择哪一个作为典型指标代替原来的两个变量, 可以根据专业知识或测定的难易程度决定。

【例 7】为更好地说明选择典型变量的计算方法, 再举一例。

有 10 个测验项目, 分别用变量  $x_1 \sim x_{10}$  表示。50 名学生参加测试。数据编号 data13-04。

要求: 对十个变量进行变量聚类; 计算并打印各变量间的相关矩阵, 用相关测度各变量间的距离。打印出聚为两类的结果即各变量属于两类中的哪一类。打印出聚类全过程的冰柱图, 以便对于变量分类进行进一步的探讨。

根据要求的操作步骤如下:

(1) 读取数据文件 data13-04。按 Analyze→Classify→Hierarchical Cluster 顺序单击菜单项, 展开分层聚类分析对话框。

(2) 在主对话框中指定分析变量, 在变量表中选择  $x_1 \sim x_{10}$ , 移到 Variables 框中。

(3) 在 Cluster 栏中选择 Variables 项, 即选择进行变量聚类。

(4) 在主对话框中按 Method 按钮, 在打开的对话框中选择聚类方法: 在 Cluster 列表中选择 Furthest Neighbor 项, 在 Measure 栏内, 选择 Interval, 在下拉列表中选择 Pearson Correlation 项, 即皮尔逊相关。

(5) 在主对话框中按 Statistics 按钮展开相应的对话框, 选择输出项: 选择 Proximity Matrix 要求打印相关矩阵。在 Cluster Membership 栏中, 选择 Single solution, 并在其后输入 2。

(6) 选择输出的统计图。在主对话框中, 按 Plots 按钮, 展开相应的对话框。在 Icicle 栏中选择 All clusters 项, 显示全过程冰柱图。

(7) 主对话框, 按 Paste 按钮, 在 Syntax 窗口中生成的 SPSS 程序如下:

```
PROXIMITIES x1 x2 x3 x4 x5 x6 x7 x8 x9 x10
/MATRIX OUT ('C:\WINDOWS\TEMP\ spss3016\ spssclus.tmp') /VIEW=VARIABLE
/MEASURE=CORRELATION /PRINT NONE /STANDARDIZE=VARIABLE NONE .
CLUSTER
/MATRIX IN ('C:\WINDOWS\TEMP\ spss3016\ spssclus.tmp')
/METHOD COMPLETE /PRINT SCHEDULE CLUSTER(2) /PRINT DISTANCE
/PLOT VICICLE. ERASE FILE='C:\WINDOWS\TEMP\ spss3016\ spssclus.tmp'.
```

在 PROXIMITIES 命令中给出参与分析的变量表。MATRAX 子命令要求计算相关矩阵并保存到指定位置 C:\WINDOWS\TEMP\spssclus.tmp, 供后续过程进行分析时使用。VIEW 子命令指定相似性计算是针对变量聚类。STANDARDIZE=NONE 命令不要求对原始数据进行标准化。

CLUSTER 命令中的 MATRIX IN 子命令要求分析 PROXIMITIES 生成的包含相关矩阵的数据文件。METHOD 子命令要求使用完全连接法即最远邻法进行聚类。PRINT 子



命令要求输出聚类过程表和分为两类的聚类结果。PLOT 子命令要求作全过程的纵向冰柱图。ERASE FILE 命令删除临时相关矩阵数据文件。

(8) 运行 Syntax 窗口中的程序或在主对话框单击 OK 按钮提交运行。在 Output 窗口中输出结果见表 13-21、表 13-22、图 13-33，其中略去了数据的综合信息。

表 13-21 变量聚类的相关系数矩阵

Proximity Matrix										
Case	Matrix File Input									
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	1.000	.133	.290	.099	.331	.198	.449	.323	.320	.112
x2	.133	1.000	.026	.411	.201	.328	.134	.199	.268	.271
x3	.290	.026	1.000	.151	.274	.406	.443	.509	.598	.318
x4	.099	.411	.151	1.000	.072	.282	.145	.401	.324	.407
x5	.331	.201	.274	.072	1.000	.317	.191	.063	.356	.084
x6	.198	.328	.406	.282	.317	1.000	.370	.312	.306	.296
x7	.449	.134	.443	.145	.191	.370	1.000	.337	.313	.246
x8	.323	.199	.509	.401	.063	.312	.337	1.000	.611	.584
x9	.320	.268	.598	.324	.356	.306	.313	.611	1.000	.325
x10	.112	.271	.318	.407	.084	.296	.246	.584	.325	1.000

表 13-22 两类的类成员

Cluster Membership	
Case	2 Clusters
x1	1
x2	1
x3	2
x4	1
x5	1
x6	1
x7	1
x8	2
x9	2
x10	2

(9) 输出结果说明

表 13-21 是测度变量间距离的相关矩阵，表中省略了各变量与本身的相关系数 1。表 13-22 是聚为两类的结果，一类是由  $x_1$ 、 $x_2$ 、 $x_4$ 、 $x_5$ 、 $x_6$ 、 $x_7$  组成，一类由  $x_3$ 、 $x_8$ 、 $x_9$ 、 $x_{10}$  组成。

图 13-33 是聚类全过程的冰柱图，可以看出分成两类的结果与表 13-22 是一致的，还可以查看如果聚为三类，各类组成为： $x_{10}$ 、 $x_9$ 、 $x_8$ 、 $x_3$ ； $x_4$ 、 $x_2$ ； $x_6$ 、 $x_5$ 、 $x_7$ 、 $x_1$ 。若聚为四类，各类组成为： $x_{10}$ 、 $x_9$ 、 $x_8$ 、 $x_3$ ； $x_2$ 、 $x_4$ ； $x_5$ 、 $x_6$ ； $x_7$ 、 $x_1$ 。

在实际工作中可以根据冰柱图和专业知识确定聚为几类最为合理，最后得出结论。

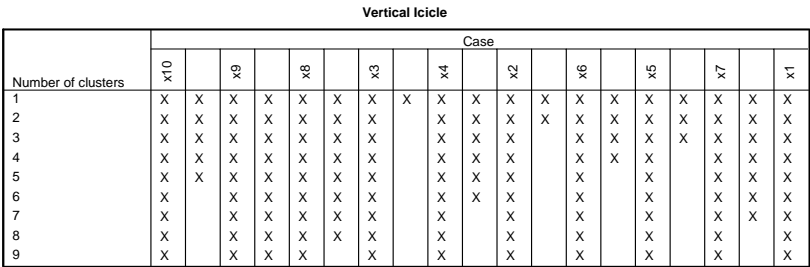


图 13-33 变量聚类全过程的冰柱图

(10) 典型指标的选择。根据下面公式计算

$$\bar{R}_j^2 = \frac{\sum r^2}{m_j - 1}$$

选择典型指标或称代表变量。以分为三类的第一组为例，计算  $\bar{R}_{10}^2$ 、 $\bar{R}_9^2$ 、 $\bar{R}_8^2$ 、 $\bar{R}_3^2$ 。方法如下：

- ① 按 Analyze→Correlate→Bivariate 顺序展开 Bivariate Correlate 相关分析对话框。
- ② 选择分析变量  $x_3$ 、 $x_8$ 、 $x_9$ 、 $x_{10}$ 。
- ③ 选择分析方法, 在 Correlation Coefficients 栏选择 Pearson。
- ④ 单击 OK 按钮提交运行, 得到表 13-23 相关矩阵表。

表 13-23 第一组变量相关矩阵

Correlations					
		x3	x8	x9	x10
x3	Pearson Correlation	1	.509**	.598**	.318*
	Sig. (2-tailed)		.000	.000	.025
	N	50	50	50	50
x8	Pearson Correlation	.509**	1	.611**	.584**
	Sig. (2-tailed)	.000		.000	.000
	N	50	50	50	50
x9	Pearson Correlation	.598**	.611**	1	.325*
	Sig. (2-tailed)	.000	.000		.021
	N	50	50	50	50
x10	Pearson Correlation	.318*	.584**	.325*	1
	Sig. (2-tailed)	.025	.000	.021	
	N	50	50	50	50

\*\*. Correlation is significant at the 0.01 level (2-tailed).

\*. Correlation is significant at the 0.05 level (2-tailed).

从表中读取相关系数, 计算各相关指数:

$$\overline{R}_3^2 = (0.509^2 + 0.598^2 + 0.318^2) / 3 = 0.23927$$

$$\overline{R}_8^2 = (0.509^2 + 0.611^2 + 0.584^2) / 3 = 0.32449$$

$$\overline{R}_9^2 = (0.598^2 + 0.611^2 + 0.325^2) / 3 = 0.27885$$

$$\overline{R}_{10}^2 = (0.318^2 + 0.584^2 + 0.325^2) / 3 = 0.18260$$

比较 4 个相关指数,  $x_8$  的相关指数最大, 因此该组变量选择  $x_8$  作代表变量。其余各组的代表变量读者可以自己按上述方法计算。

### 13.4.6 分层聚类过程的命令语句

分层聚类分析涉及两个 SPSS 过程: PROXIMITIES 和 CLUSTER。下面分别介绍。

#### 1. PROXIMITIES 过程语句

PROXIMITIES varlist [/VIEW={CASE\*\*}]{VARIABLE}}

[/STANDARDIZE={[VARIABLE]{CASE}}][{NONE\*\*}{Z}{SD}{RANGE}{MAX}  
{MEAN}{RESCALE}]]

[/MEASURE={[EUCLID\*\*]{SEUCLID}{COSINE}{CORRELATION}{BLOCK}  
{CHEBYCHEV}{POWER(p,r)}{MINKOWSKI(p)}{CHISQ}{PH2}  
{RR}([p],[np]]){SM}([p],[np]])}{JACCARD}([p],[np]]){DICE}([p],[np]])  
{SS1}([p],[np]]){RT}([p],[np]]){SS2}([p],[np]]){K1}([p],[np]])  
{SS3}([p],[np]]){K2}([p],[np]]){SS4}([p],[np]]){SS5}([p],[np]])  
{OCHIAI}([p],[np]]){HAMANN}([p],[np]]){PHI}([p],[np]])}]

```

{LAMBDA[(p[,np])]}{D[(p[,np])]}{Y[(p[,np])]}{Q[(p[,np])]}
{BEUCLID[(p[,np])]}{SIZE[(p[,np])]}{PATTERN[(p[,np])]}
{BSEUCLID[(p[,np])]}{BSHAPE[(p[,np])]}{DISPER[(p[,np])]}
{VARIANCE[(p[,np])]}{BLWMN[(p[,np])]}{NONE}}
[ABSOLUTE] [REVERSE] [RESCALE]
[/PRINT={[PROXIMITIES**] {NONE} }] [/ID=varname]
[/MISSING=[LISTWISE**] [INCLUDE] [/MATRIX=[IN({file})* ]] [OUT({file})*]]

```

**PROXIMITIES** 是命令语句关键字, 其后的变量表是要用该命令处理的变量。

“[ ]”中的均为子命令, 一个子命令中的若干个并列的“[ ]”中的内容表明该子命令完成其功能需要指定的参数。是子命令下属的第一层选项, 并列的几个选项可以取其一, 也可以取其中的若干项。在这一层选项下面并列的若干个用“{ }”包含的内容是第二层选项, 并列的第二层选项中只能取其一。

## 2. 子命令及其含义

(1) **VIEW** 子命令指定近似计算的目的, 有两种选项: **CASE** 是默认值, 表明是要进行观测量聚类; **VARIABLE** 表明是要进行变量聚类。

(2) **STANDARDIZE** 子命令对变量或观测量进行标准化处理。有两个并列的选项:

① 选择标准化对象: **VARIABLE** 对变量进行标准化; **CASE** 对观测量进行标准化。

② 选择对数值进行标准化的算法: **NONE** 不进行标准化; **Z** 标准化到 **Z** 分数 (均值为 0, 标准差为 1); **SD** 标准化到标准差为 1; **RANGE** 标准化到使所有值的范围均在 -1~+1 之间; **MAX** 标准化到最大值为 1; **MEAN** 标准化到均值为 1。 **RESCALE** 标准化到使所有值均在 0~1 范围内。有关算法请见附录。

(3) **MEASURE** 子命令指定对距离测度的方法。这里所说的“距离”是广义的距离, 在聚类分析中实际上是对相似性 (或不相似性) 的测度。无论是相似性还是实际上的两者之间的距离在这里统统称为“距离”。该子命令后面是选择的计算距离或相似性、不相似性测度的方法名。对有的测度方法需要在后面的括号中指定参数。

对二值数据, 后面括号中的参数表示某特性是否出现的数值标志, 如果出现两个参数, 第一个参数代表特性出现的数值标志, 第二个参数表示特性不出现的标志。如果使用一个参数, 该参数表示特性出现的数值标志, 其他值, 一律认为是特性不出现的标志。如果没出现参数, 则使用默认值: 1 表示特性出现, 0 表示特性不出现。例如子命令:

**Measure=RR(1,2)**表示使用匹配系数 **RR** 测度两个二值变量或观测量之间的距离。1 代表某特性出现, 2 代表某特性不出现。 **CLUSTER** 过程根据这一距离矩阵进行聚类。

可以选择的测度距离的方法有多种, 下面举例说明:

① 对连续变量的距离 (或不相似性) 测度方法有以下八种: **EUCLID**、**SEUCLID**、**COSIN**、**CORR**、**BLOCK**、**CHEBYSCHER**、**MINKOSKI(p)**、**POWER(p,r)**。具体含义和算法见附录。

② 对于计数变量的距离或不相似性的测度方法共两个: CHISQ 是卡方、 $PH^2$  即  $\phi^2$ 。

③ 对二值变量的距离或不相似性的测度方法有 27 个。语句中的各种算法的关键字基本是各方法英文单词的字头。计算方法可参见附录 A 有关内容。

④ 对测度的变换方法: ABSOLUTE、REVERSE、RESCALE 具体含义见附录。

(4) PRINT 子命令把根据 METHOD 子命令、STANDARDIZE 子命令和 MEASURE 子命令对原始数据进行的近似计算和数值转换的结果输出到输出观察窗口。有两种选择:

① PROXIMITIES, 输出近似计算的结果, 此为系统默认的选择。

② NONE, 不输出近似计算的结果。

(5) ID 子命令指定标识观测量的变量名。

(6) MISSING 子命令指定对缺失值的处理方法, 有两种可供选择的处理方法:

① LISTWISE, 剔除变量表中变量带有缺失值的观测量。是默认的处理方法。

② NCLUDE, 不剔除带有读者定义的缺失值的观测量。

(7) MATRIX 子命令

PROXIMITIES 过程使用数据窗口中的数据文件作为输入数据文件进行近似计算。否则需要用 MATRIX 子命令指定输入数据文件。PROXIMITIES 过程的计算结果往往作为其他分析过程的输入数据。因此需要使用 MATRIX 子命令指定输出数据文件, 以便需要时在分析过程中用一个类似的子命令读入这些数据。MATRIX 子命令中的两个选项可以单独使用, 也可以同时使用。

① IN, 指定输入数据文件名。

② OUT, 指定输出数据文件名, 文件名放在后面的括号中。必要时应该同时指定文件的路径。如果在应该指定输出文件名处使用了星花 “\*”, 该距离矩阵不输出到磁盘上, 而是在数据窗口中代替当前的工作数据文件。

### 3. CLUSTER 过程的命令语句

#### (1) 命令格式

CLUSTER varlist [/MISSING=LISTWISE\*\*] [INCLUDE]

[/MEASURE={SEUCLID\*\*} {EUCLID} {COSINE}]

{CORRELATION} {BLOCK} {CHEBYCHEV} {POWER(p,r)} {MINKOWSKI(p)}

{CHISQ} {PH2} {RR[(p[,np])]} {SM[(p[,np])]} {JACCARD[(p[,np])]} {

DICE[(p[,np])]} {SS1[(p[,np])]} {RT[(p[,np])]} {SS2[(p[,np])]} {

K1[(p[,np])]} {SS3[(p[,np])]} {K2[(p[,np])]} {SS4[(p[,np])]} {HAMANN[(p[,np])]} {

OCHIAI[(p[,np])]} {SS5[(p[,np])]} {PHI[(p[,np])]} {LAMBDA[(p[,np])]} {D[(p[,np])]} {

Y[(p[,np])]} {Q[(p[,np])]} {BEUCLID[(p[,np])]} {SIZE[(p[,np])]} {PATTERN[(p[,np])]} {

BSEUCLID[(p[,np])]} {BSHAPE[(p[,np])]} {DISPER[(p[,np])]} {VARIANCE[(p[,np])]} {

BLWMN[(p[,np])]

[/METHOD={BAVERAGE\*\*} {WAVERAGE} {SINGLE} {COMPLETE} {CENTROID}]

```
[MEDIAN ] {WARD } (DEFAULT** ) [(rootname)] [...]  
[SAVE=CLUSTER({level } {min,max})] [/ID=varname]  
[/PRINT=[CLUSTER({level } {min,max})] [DISTANCE] [SCHEDULE**] [NONE]]  
[/PLOT=[VICICLE**[(min [, max [, Inc]])] [DENDROGRAM] [NONE]]  
[HICICLE [(min [, max [, Inc]])] ] [MATRIX=[IN({file } {* })] [OUT({file } {*})]]
```

其中: **CLUSTER** 是命令关键字, 其后的变量表是要用该命令处理的变量。如果使用 **MATRIX** 子命令及其 **IN** 子命令, 说明原始数据已经经过近似计算, 中间结果在 **IN** 子命令指定的数据文件中, 那么应该在命令关键字 **CLUSTER** 后面不出现变量表。

(2) 子命令及其含义。全部子命令均为可选择的子命令。

① **MISSING** 子命令, 该子命令指定分析过程中对缺失值的处理方法, 系统默认分析中剔除系统缺失值和读者定义的缺失值, 读者可以选择下面两个选项:

- **LISTWISE**, 在分析变量表中的变量遇到带有缺失值的观测量 (或变量) 不再计算该观测量 (或变量) 与其他观测量 (或变量) 间的距离。此选项是系统默认值。
- **INCLUDE**, 不剔除带有读者定义的缺失值的观测量或变量。但应该清楚: 系统默认的是将带有系统缺失值和读者缺失值的观测量或变量从分析中剔除。

② **MEASURE** 子命令与 **PROXIMITIES** 过程的 **MEASURE** 子命令的选项基本一致。

③ **METHOD** 子命令指定进行聚类分析的方法, 有 7 种可供选用的方法: **BAVERAGE** 组间连接法, 是系统默认的聚类方法、**WAVERAGE** 组内连接法、**SINGLE** 最近邻法、**COMPLETE** 最远邻法或称完全连接法、**CENTROID** 重心法、**MEDIAN** 中位数法、**WARD** 沃尔德法。解释见附录。

④ **SAVE** 子命令。在输入数据文件中建立表明聚类结果的新变量。只有在进行观测量聚类时可以使用。选项只有一个, 但选项中的参数有两种:

- **CLUSTER( $n$ )**, 要求新变量表明聚为  $n$  类的聚类结果, 即表明每个观测量分属到哪一类。其值为  $1 \sim n$ ,  $n$  为正整数。
- **CLUSTER( $n_1, n_2$ )**,  $n_1, n_2$  均为正整数, 且  $n_2 > n_1$ 。要求建立  $n_2 - n_1 + 1$  个新变量, 分别表明把观测量聚为  $n_1, n_1 + 1, n_1 + 2, \dots, n_2$  类时每个观测量分派到的类号。

⑤ **ID** 子命令指定标识观测量的变量名。

⑥ **PRINT** 子命令指定输出到 Output 窗口中的内容。共有四个选项:

- **CLUSTER**, 有两种形式: 与 **SAVE** 子命令中选项 **CLUSTER** 的两种形式对应。
- **CLUSTER( $n$ )**, 输出聚为  $n$  类时的聚类结果表。
- **CLUSTER( $n_1, n_2$ )**,  $n_1, n_2$  均为正整数, 且  $n_2 > n_1$ 。该项要求在输出窗口中显示聚类结果表, 表明观测量被聚为  $n_1, n_1 + 1, n_1 + 2, \dots, n_2$  类时每个观测量被分派到的类号。
- **DISTANCE**, 输出各变量或观测量之间的距离数据。
- **SCHEDULE**, 输出反映聚类过程的凝聚状态表。此为系统默认选项。
- **NONE**, 不输出任何信息到 Viewer 窗口。

⑦ PLOT 子命令指定要求输出的统计图。共三种：纵向、横向冰柱图和树形图。

- VICICLE[( $n_1$ [, $n_2$ [, $m$ ]])], 纵向冰柱图, 是系统默认的。选项有下列四种形式:

VICICLE, 后面不指定任何参数, 纵向冰柱图表明聚类全过程。

VICICLE( $n_1$ ), 括号中有一个参数, 要求冰柱图表明聚为  $n_1$  类的每一步过程。

VICICLE( $n_1, n_2$ ), 要求冰柱图表明聚为  $n_1$  类到聚为  $n_2$  类的每一步聚类过程。

VICICLE( $n_1, n_2, m$ ), 要求冰柱图表明聚为  $n_1$  类到聚为  $n_2$  类的间隔为  $m$  步的聚类过程。

选项中的参数  $n_1$ 、 $n_2$ 、 $m$  均为正整数。

- HICICLE [( $n_1$ [, $n_2$ [, $m$ ]])], 横向冰柱图选项有四种形式。括号中参数的规定与纵向冰柱图参数规定相同。

- DENTROGRAM, 要求输出反映聚类全过程的树形图。

- NONE, 要求不输出任何统计图。

⑧ MATRIX 子命令指定 CLUSTER 过程使用的输入和输出数据文件。它有两个选项:

IN, 该选项指定用于进行聚类分析的数据文件, 一般情况下使用 PROXIMITIES 过程近似计算的输出文件, 即与 PROXIMITIES 过程的 MATRIX 子命令中的 OUT 选项指定的数据文件对应。

OUT, 为存储矩阵数据文件指定一个带有存储路径的文件名。如果使用星号, 则表示将矩阵数据文件置于工作数据窗口中, 代替当前的工作数据文件。

## 13.5 判 别 分 析

### 13.5.1 判别分析概述

#### 1. 判别分析的概念

判别分析是一种常用的统计分析方法。判别分析是根据观察或测量到若干变量值, 判断研究对象属于哪一类的方法。例如医学实践中根据各种化验结果、疾病症状、体征判断患者患的是什么疾病。体育人才选拔是根据运动员的体形、运动成绩、生理指标、心理素质指标、遗传因素判断是否选入运动队继续培养。动物、植物分类等都可以用判别分析来解决。判别分析是应用计算机进行运动员选才、动物、植物分类以及疾病辅助诊断等的主要统计学基础。

进行判别分析必须已知观测对象的分类和若干表明观测对象特征的变量值。判别分析就是要从中筛选出能提供较多信息的变量并建立判别函数, 使得利用推导出的判别函数对观测量判别其所属类别时的错判率最小。

线性判别函数一般形式是:  $y = a_1x_1 + a_2x_2 + a_3x_3 + \Lambda + a_nx_n$

其中:  $y$  为判别分数 (判别值),  $x_1, x_2, x_3, \dots, x_n$  为反映研究对象特征的变量,  $a_1, a_2, a_3, \dots, a_n$  为各变量的系数, 也称判别系数。

SPSS 对于分为  $m$  类的研究对象,建立  $m$  个线性判别函数。对于每个个体进行判别时,把测试的各变量值代入判别函数,得出判别分数,或者计算属于各类的概率,从而确定该个体属于哪一类。还建立标准化和未标准化的典则判别函数。

## 2. Discriminant 过程的功能

SPSS 提供的判别分析过程是 Discriminant 过程。它根据已知的观测量分类和表明观测量特征的变量值推导出判别函数,并把各观测量的自变量值回代到判别函数中,根据判别函数对观测量所属类别进行判别。对比原始数据的分类和按判别函数所判的分类,给出错分概率。

判别分析可以根据类间协方差矩阵,也可以根据类内协方差矩阵进行分析。如果原始数据中的观测量是分为  $m$  类的,每一已知类的先验概率可以取其值相等即等于  $1/m$ ,也可以与各类样本量成正比。

判别分析可以根据要求,给出各类观测量的单变量的描述统计量、线性(费雪 Fisher)判别函数的系数或标准化及未标准化的典则判别函数的系数、类内相关矩阵、类内和类间协方差矩阵和总协方差矩阵,给出按判别函数判别(回代)的各观测量所属类别,带有错分率的判别分析小结,还可以根据要求生成表明各类分布的区域图和散点图。如果希望把部分聚类结果存入文件,还可以在工作数据文件中建立新变量,表明观测量按判别函数分派的类别、按判别函数计算的判别分数和分到各类去的概率。

Discriminant 过程的大部分功能都可以通过对话框来指定,还有一些功能可以在 Syntax 窗口中给予补充或修改。例如指定各类的先验概率、显示旋转方式和结构矩阵、限制提取的判别函数的数目、读取一个相关矩阵、分析后把相关矩阵写入文件、指定对参与分析的观测量进行回代分类,对没有参与分析的观测量进行预测分类等。这部分可以查看 13.5.4 节。

## 3. 有关判别分析的术语

(1) 建立判别函数的方法有 4 种:全模型法、向前选择法、向后选择法、逐步选择法。

① 全模型法。全模型法是把读者指定的变量全部放入判别函数中,不管变量对判别函数是否起作用,作用大小如何。当对反映研究对象特征的变量认识比较全面时可以选择此种方法。此种方法是 SPSS 系统默认的方法。

由于人们对客观事物的认识可能并不客观,因此对变量的选择就有可能出现偏差。如果没有选择对研究对象的特征能够提供丰富信息的变量,没有测试有关的数据,只能等待人们对所研究事物认识的进一步深化,别无他法。但是,如果选择的变量中有对研究对象的特征不能提供较丰富的信息,对判别贡献很小的变量,这样的变量就应该从判别模型中剔除。全模型方法不能解决这个问题。

② 向前选择法。该方法是从判别模型中没有变量开始,每一步把一个对判别模型的判断能力贡献最大的变量引入模型,直到在没有被引入模型的变量中没有符合进

入模型的条件（判据）时，变量引入过程结束。当希望比较多的变量留在判别函数中时，使用向前选择法。

③ 向后选择法。此方法与向前选择法完全相反，它是从把读者所有指定的变量建立一个全模型，每一步把一个对模型的判断能力贡献最小的变量剔除出模型，直到模型中的所有变量都符合留在模型中的判据时，剔除变量工作结束。在希望较少的变量留在判别函数中时使用向后选择法。

④ 逐步选择法。此判别法从模型中没有变量开始，每一步都要对模型进行检验。每一步都在把模型外的对模型的判别能力贡献最大的变量加入到模型中的同时，也考虑把已经在模型中但又不符合留在模型中的条件的变量剔除。这是因为新变量的引入有可能使原来已经在模型中的变量对模型的贡献变得不显著了。直到模型中的所有变量都符合引入模型的判据，模型外的变量都不符合进入模型的判据时，逐步选择变量的过程停止。逐步选择法更能比较好地选择变量，SPSS 用此种方法建立非全（变量）判别函数。此种方法作为可选择的方法。

(2) 典则判别分析。典则判别分析建立典则变量代替原始数据文件中指定的自变量。典则变量是原始自变量的线性组合。用少量的典则变量代替原始的多个变量可以比较方便地描述各类之间的关系，例如可以用平面区域图或散点图直观地表示各类之间的相对关系。SPSS 计算标准化和未标准化的典则判别函数系数。

(3) 判别函数的性能。判别分析得出的判别函数性能如何，可以通过回代的方法进行验证。即将各观测量的变量值代到线性判别函数中，根据线性判别函数值（判别分数）确定每个观测量分属于哪一类，然后与原始数据中的分类变量值进行比较，得到错判率。错判率越小说明判别函数的判别性能越好。

(4) 判别分析对数据的要求。进行判别分析要求数据遵循多元正态分布。实践工作中收集的数据，其分布往往不同于正态分布，因此使用本节介绍的参数分析方法是不合适的。从非正态总体导出的线性判别函数（或经过预处理的数据）导出的二次判别函数的误差率估计可能会有较大的偏差。

#### (5) 利用判别函数对观测量进行分类

用 Discriminant 过程导出的线性判别函数的数目与类别数目相同。确定一个观测量属于哪一类，可以把该观测量的各变量值代入每个判别函数，哪个判别函数值大，该观测量就属于哪一类。

### 13.5.2 判别分析过程

#### 1. 建立或读入数据文件

在数据窗口中输入待分析的数据或利用 File 菜单中的 Open 命令打开已经存在的数据文件，显示到数据窗口中。数据中必须包括一个表明已知的观测量所属类别的变量和若干个表明分类特征的变量。



2. 按 **Analyze→Classify→Discriminat** 顺序单击菜单项, 展开 **Discriminant Analysis** 判别分析的主对话框, 见图 13-34。

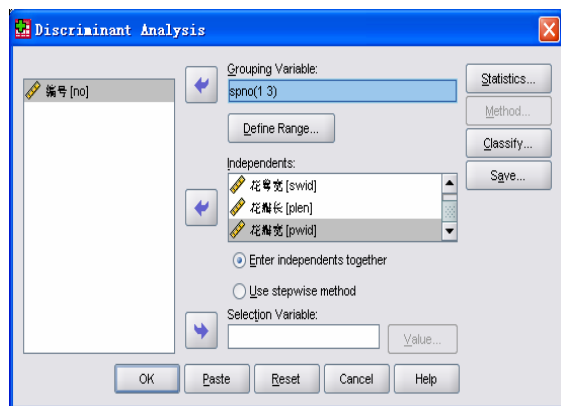


图 13-34 判别分析的主对话框

### 3. 选择分类变量及其范围

在主对话框中源变量表中选择表明已知的观测量所属类别的变量（一定是离散变量）送入 **Grouping Variable** 框中。此时矩形框下面的 **Define Range** 按钮加亮, 单击该按钮, 打开定义分类变量范围的小对话框, 如图 13-35 所示。在 **Minimum** 框中输入该分类变量的最小值, 在 **Maximum** 框中输入该分类变量的最大值。

### 4. 指定判别分析的自变量

在主对话框的源变量表中选择表明观测量特征的变量, 送到 **Independents** 框中, 作为参与判别分析的变量。

5. 完成前面 4 步骤的操作, 即可使用系统默认值对工作数据集的数据进行判别分析了。单击 **OK** 按钮提交执行, 在输出窗口中显示出分析结果。

完全使用系统默认值进行判别分析, 结果有时不能令人满意, 因此根据以下步骤指定选项是很有必要的。

### 6. 选择观测量

如果希望使用一部分观测量进行判别函数的推导, 而且有一个变量的某个值可以作为这些观测量的标识, 则可以在主对话框中。从源变量表框中选择这个变量送入 **Selection variable** 栏中, 再单击 **Value** 按钮, 展开 **Set Value** 子对话框如图 13-36 所示, 输入标识参与分析的观测量所具有的所有合法观测量。此步骤可以省略。

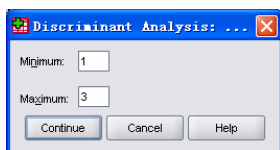


图 13-35 指定分类变量范围

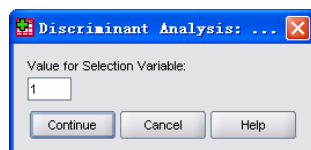


图 13-36 设置选择观测量值

### 7. 选择分析方法

在主对话框中自变量矩形框下面有两个选项, 可从中选择判别分析方法。

(1) **Enter independent together**. 当认为所有自变量都能对观测量的特性提供丰富的信息, 且彼此独立时使用该选项。判别分析过程将不加选择地使用所有自变量进行判别分析, 建立全模型, 不需要进一步进行选择。

(2) **Use stepwise method**. 当不认为所有自变量都能对观测量的特性提供丰富的信息时, 使用该选项。

单击 Method 按钮, 展开 Stepwise Method 对话框, 见图 13-37。

① 在 Method 栏可供选择的判别分析方法有:

- Wilks' lambda, 每步都是 Wilk 的  $\lambda$  统计量最小的进入判别函数。
- Unexplained variance, 每步都是各类不可解释的方差和最小的变量进入判别函数。
- Mahalanobis distance, 每步都使靠得最近的两类间的 Mahalanobis 距离最大的变量进入判别函数。

• Smallest F ratio, 每步都是使任何两类间最小的  $F$  值最大的变量进入判别函数。

• Rao's V, 每步都是使 Rao's V 统计量产生最大增量的变量进入判别函数。可以对一个要加入到模型中的变量的  $V$  值指定一个最小增量。选择此种方法后, 应该在该项下面的 V-to-enter 后的矩形框中输入这个增量的指定值。当某变量导致的  $V$  值增量大于指定值的变量进入判别函数。

② 选择逐步判别停止的判据在 Criteria 栏中进行, 可供选择的判据有:

• Use F value, 使用  $F$  值, 是系统默认的判据。当加入一个变量 (或剔除一个变量) 后, 对在判别函数中的变量进行方差分析。当计算的  $F$  值大于指定的 Entry 值时, 该变量保留在函数中, 默认值是 Entry 为 3.84。当该变量使计算的  $F$  值小于指定的 Removal 值时, 该变量从函数中剔除, 默认值是 Removal 为 2.71。设置这两个值时应该注意使 Entry > Removal, 否则产生函数中没有变量的错误。

• Use probability of F, 用  $F$  检验的概率决定变量是否加入函数或被剔除。加入变量的  $F$  值概率的默认值是 0.05 (5%), 移出变量的  $F$  值概率是 0.10 (10%)。Removal 值应该 (移出变量的  $F$  值概率) > Entry 值 (加入变量的  $F$  值概率)。

③ 显示内容的选择。Display 栏中的两项选择要显示的统计量:

- Summary of steps, 要求在逐步选择过程中的每一步之后显示每个变量的统计量。
- F for Pairwise distances, 要求显示两两类之间的两两  $F$  值矩阵。

## 8. 指定输出的统计量

单击 Statistics 按钮, 展开 Statistics 子对话框, 如图 13-38 所示。

(1) 在 Descriptives 栏中选择要输出的原始数据的描述统计量:

① Means, 输出各类中各自变量均值 MEAN、标准差 Std Dev 和各自变量总样本的均值和标准差。

② Univariate ANOVAs, 要求进行假设检验, 输出单变量方差分析结果。检验的无效假设是: 各类中同一自变量均值都相等。

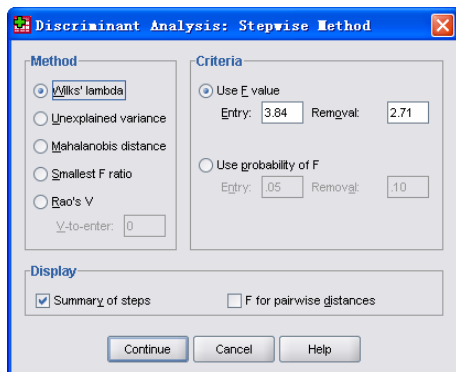


图 13-37 判别分析方法选择对话框

③ Box's M, 对各类的协方差矩阵相等的假设进行检验。如果样本足够大, 表明差异不显著的  $p$  值意味着矩阵差异不明显。

(2) 在 Function coefficients 栏中选择判别函数系数的输出形式:

- Fisher's, 要求输出可以直接用于对新样本进行判别分类的费雪系数。对每一类给出一组系数, 并给出该组中判别分数最大的观测量。
- Unstandardized, 输出未经标准化的判别系数。

(3) 在 Matrices 栏中选择要求给出的自变量系数矩阵:

- Within-groups correlation, 要求输出类内相关矩阵, 它是根据计算相关矩阵之前, 将各组 (类) 协方差矩阵平均后计算的。
- Within-groups covariance, 要求计算并显示合并类内协方差矩阵, 是将各组 (类) 协方差矩阵平均后计算的, 区别于总协方差矩阵。
- Separate-groups covariance, 对每类输出显示一个协方差矩阵。
- Total covariance, 计算并显示总样本的协方差矩阵。

9. 指定分类参数和判别结果

在主对话框中单击 Classify 按钮, 展开相应的对话框, 如图 13-39 所示。

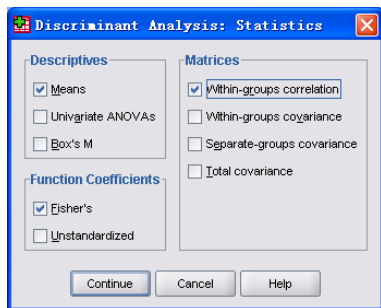


图 13-38 选择输出统计量的对话框

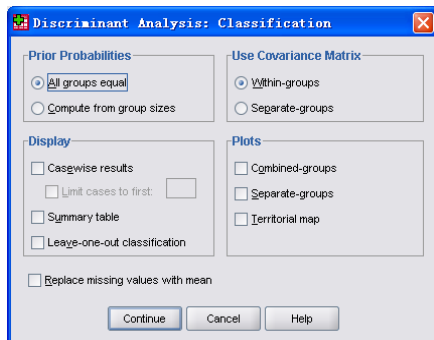


图 13-39 选择参数与结果的对话框

(1) 在 Prior Probabilities 栏中选择先验概率:

- All groups equal, 各类先验概率相等。若分为  $m$  类, 则各类先验概率均为  $1/m$ 。
- Compute from group sizes, 各类的先验概率与各类的样本量成正比。

(2) 在 Use Covariance Matrix 栏中选择分类使用的协方差矩阵。

- Within-groups, 指定使用合并组内协方差矩阵进行分析。
- Separate-groups, 指定使用各组协方差矩阵进行分析。

(3) 在 Display 栏中选择生成到输出窗口中的分类结果, 选项有:

- Casewise results, 对每个观测量输出判别分数、实际类、预测类 (根据判别函数求得的分类结果) 和后验概率等。选择此项, 还可以选择其附属选项 Limits cases to First,

并在后面的小矩形框中输入观测量数  $n$ ，含义为仅输出前  $n$  个观测量的分类结果。观测数量大时可以选择此项。

- **Summary table**，输出分类小结，给出正确分类观测量数，即原始类和根据判别函数计算的预测类相同的观测量数、错分观测量数和错分率。

- **Leave-one-out classification**，输出每个观测量的分类结果，所依据的判别函数是由除该观测量以外的其他观测量导出的，因此也称为交互校验结果。

(4) 在 **Plots** 栏中选择要求输出的统计图。可以同时选择几种输出的统计图形：

- **Combined-groups**，生成一张包括各类的散点图。该散点图是根据前两个判别函数值做出的。如果只有一个判别函数，就输出直方图。

- **Separate-groups**，根据前两个判别函数值对每一类生成一张散点图，共分为几类就生成几张散点图。如果只有一个判别函数，就输出直方图。

- **Territorial map**，根据函数值生成把观测量分到各组中去的区域图。此种统计图把一张图的平面划分出与类数相同的区域，每一类占据一个区，各类的均值在各区中用星号标出。如果仅有一个判别函数，则不作此图。

(5) 缺失值处理方式在 **Classification** 子对话框的最下面选项，选中 **Replace missing value with mean**，则用该变量的均值代替缺失值。

#### 10. 指定生成并保存在数据文件中的新变量

**Discriminant** 过程可以在数据文件中建立新变量。在主对话框中单击 **Save** 按钮，展开如图 13-40 所示的子对话框。

① **Predicted group membership** 要求建立一个新变量，其值是根据判别分数、按后验概率最大预测的分类。每运行一次 **Discriminant** 过程，就建立一个表明使用判别函数预测的各观测量属于哪一类的新变量。第一次运行建立新变量的变量名为 **dis\_1**，如果在工作数据文件中不把前一次建立的新变量删除，第  $n$  次运行建立的新变量默认变量名为 **dis\_1**。

② **Discriminant score**，建立表明判别分数的新变量。该分数是由未标准化的判别系数乘自变量的值，将这些乘积求和后加上常数得来。每次运行 **Discriminant** 过程都给出一组表明判别分数的新变量。建立几个判别函数就有几个判别分数变量。参与分析的观测量共分为  $m$  类，则建立  $m-1$  个典则判别函数，指定该选项，就可以生成  $m-1$  个表明判别分数的新变量。例如原始数据观测量共分为 3 类，建立两个典则判别函数。第一次运行判别过程建立的新变量名为 **dis1\_1**、**dis2\_1**，第二次运行判别过程建立的新变量名为 **dis1\_2**、**dis2\_2**、…以此类推。分别表示代入第一和第二个判别函数所得到的判别分数。

③ **Probabilities of group membership**，要求建立新变量表明观测量属于某一类的概

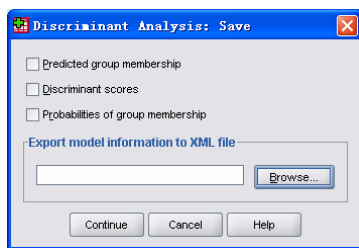


图 13-40 建立新变量对话框

率。有  $m$  类，对一个观测量就会给出  $m$  个概率值，因此建立  $m$  个新变量。例如原始和预测分类数是 3，指定该选项，在第一次运行判别过程后，给出的表明分类概率的新变量名为 dis1\_2、dis2\_2、dis3\_2。

11. 在主对话框中单击 OK 按钮提交执行。

### 13.5.3 判别分析实例

【例 8】下面是统计学常用的实例，三种鸢尾花的花瓣、花萼的长、宽数据。共收集了三种鸢尾花，每种 50 个观测量，共 150 个观测量的数据。数据编号 data13-05。

在数据窗口中定义 5 个变量：slen（花萼长）、swid（花萼宽）、plen（花瓣长）、pwid（花瓣宽）是表明观测量（鸢尾花）特征的变量。spno（分类号）。分类的值标签是：1 刚毛鸢尾花（Setosa）、2 变色鸢尾花（Versicolor）、3 弗吉尼亚鸢尾花（Virginica）输入这些变量的值。观测量标识变量 no 是为核对方便设置的，非分析所需要。

1. 使用系统默认值进行分析

(1) 操作步骤

① 读取数据文件 data13-05。按 Analyze→Classify→Discriminant 顺序单击菜单项，展开 Discriminant Analysis 对话框。

② 在主对话框中进行以下操作：

在源变量栏里选择 slen、swid、plen、pwid，并移到 Independents 框中，作为自变量。

在左面的变量栏里选择变量 spno，并移到 Grouping Variable 框中，作为分类变量，单击 Define Range 按钮。在 Define Range 对话框中，输入变量 spno 的数值范围，Minimum 框中输入 1，Maximum 框中输入 3。

③ 单击 OK 按钮，提交系统执行。运行的命令语句是：

DISCRIMINANT	①
/GROUPS=spno(1 3)	②
/VARIABLES=slen swid plen pwid	③
/ANALYSIS ALL	④
/PRIORS EQUAL	⑤
/CLASSIFY=NONMISSING POOLED.	⑥

(2) 程序解释

① DISCRIMINANT 是判别分析过程的过程名（关键字）。“/”后面的是子命令。

② /GROUPS=spno(1 3) 分类子命令给出分类变量是 spno，分类值范围是 1~3。

③ /VARIABLES=slen swid plen pwid 变量子命令列出自变量表。

④ /ANALYSIS ALL 分析子命令是由于 Select 选项使用默认值的结果，说明无选择地使用所有观测量进行判别分析，建立全变量模型。

⑤ /PRIORS EQUAL 子命令是 Classify 选项的 Prio Probabilities 使用默认值的结果，

指定各类先验概率相等，是 0.3333。

⑥ /CLASSIFY=NONMISSING POOLED 子命令是 Classify 选项的 Use Covariance Matrixs 使用系统默认值 Within-groups 的结果，指定分类时使用合并类内协方差矩阵。无缺失值，因此对缺失值不进行处理。

(3) 输出结果见表 13-24~表 13-29。

表 13-24 基本数据信息

输出结果解释如下：

表 13-24 为基本数据信息：按变量 spno 确定分组、变量 spno 的标签是分类。下面同列单元格中是表明分类的 spno 变量的 3 个值标签。可以看出总共处理了 150 个（未加权）的观测量。每类中各变量都有 50 个未加权的观测量，以下的分析中将使用 150 个（未加权）的观测量。分组的观测量数据表中的数据表明 spno=1 为刚毛鸢尾花，spno=2 的是变色鸢尾花，spno=3 的是弗吉尼亚鸢尾花，各有 50 个观测量。三种鸢尾花的每个观测量的权重均为 1，总权重均为 50。共有 150 个观测量，总权重为 150。

Group Statistics			
分类		Valid N (listwise)	
		Unweighted	Weighted
刚毛鸢尾花	花萼长	50	50.000
	花萼宽	50	50.000
	花瓣长	50	50.000
	花瓣宽	50	50.000
变色鸢尾花	花萼长	50	50.000
	花萼宽	50	50.000
	花瓣长	50	50.000
	花瓣宽	50	50.000
弗吉尼亚鸢尾花	花萼长	50	50.000
	花萼宽	50	50.000
	花瓣长	50	50.000
	花瓣宽	50	50.000
Total	花萼长	150	150.000
	花萼宽	150	150.000
	花瓣长	150	150.000
	花瓣宽	150	150.000

表 13-25 给出了典则判别函数征值 Eigenvalues。给出的统计量自左至右有：Fuction 下面的单元格中的数字是函数代号；Eigenvalue 用于分析的前两个典则判别函数的特征值，是组间平方和与组内平方和之比。最大特征值与组均值最大的向量对应，第二大特征值对应着次大的组均值向量；% of Variance 是方差的百分比；Cumulative 是累计百分比，方差累计百分比最后累计值是 100%；Canonical Correlation 是典则相关系数。是组间平方和与总平方和之比的平方根。被平方的是由组间差异解释的总变异的比值。

表 13-26 为 Wilks' Lambda 统计量，该统计量进行检验的零假设是各组各变量均数相等。 $P < 0.001$  原假设成立的概率极小。说明该判别函数能将两类很好地区分开。表中自左至右各列：比较的函数编号；Wilks' Lambda 统计量值（也有称 U 统计量）值范围 0~1，越大表示组均值差异越小，值为 1 各组均值相等；Chi-square 是对 Wilks' lambda 的卡方转换，用于确定其显著性；df 用于计算显著性水平的自由度；最后一列 Sig 是假设检验成立的概率两个函数的 Sig 都很小，说明判别函数具有统计显著性。

表 13-27 是标准化典则判别函数系数表。由此表可以看出使用变量标签的两个判别函数分别如下。为分析方便，判别函数中使用的不是原变量名，而是变量标签：

$$y_1 = -0.346 \times \text{花萼长} - 0.525 \times \text{花萼宽} + 0.846 \times \text{花瓣长} + 0.613 \times \text{花瓣宽}$$

$$y_2 = 0.039 \times \text{花萼长} + 0.742 \times \text{花萼宽} - 0.386 \times \text{花瓣长} + 0.555 \times \text{花瓣宽}$$

注意，上述是标准化的典则判别函数，若要计算标准化典则判别函数值（即标准化典则判别分数），代入上述函数的自变量值必须是标准化以后的值。

表 13-28 为 Structure Matrix 结构阵，即合并类内相关阵。是判别变量与标准化的典

则判别函数之间的相关。变量按函数内相关的绝对值大小排列，每个变量和任何一个判别函数之间相关系数绝对值最大的标有“\*”。

表 13-25 典则判别函数特征值表

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	30.419 <sup>a</sup>	99.0	99.0	.984
2	.293 <sup>a</sup>	1.0	100.0	.476

a. First 2 canonical discriminant functions were used in the analysis.

表 13-26 判别函数的有效性检验

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.025	538.950	8	.000
2	.774	37.351	3	.000

表 13-27 标准典则判别函数的系数

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
花萼长	-.346	.039
花萼宽	-.525	.742
花瓣长	.846	-.386
花瓣宽	.613	.555

表 13-28 结构矩阵

	Function	
	1	2
花瓣长	.726*	.165
花萼宽	-.121	.879*
花瓣宽	.651	.718*
花萼长	.221	.340*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions. Variables ordered by absolute size of correlation within function.

\*. Largest absolute correlation between each variable and any discriminant function

表 13-29 Functions at Group Centroids 组（类）均值（重心）处的典则判别函数值。

表 13-29 类中心

Functions at Group Centroids

分类	Function	
	1	2
刚毛鸢尾花	-7.392	.219
变色鸢尾花	1.763	-.737
弗吉尼亚鸢尾花	5.629	.518

Unstandardized canonical discriminant functions evaluated at group means

刚毛鸢尾花类中心的函数值为  $y_1 = -7.392$ ,  $y_2 = 0.219$   
变色鸢尾花类中心的函数值为  $y_1 = 1.763$ ,  $y_2 = -0.737$   
弗吉尼亚鸢尾花类中心的函数值为  $y_1 = 5.629$ ,  $y_2 = 0.518$   
未标准化的典则判别函数中心值在各变量均值处。

【例 9】仍然使用鸢尾花的实例说明选项的作用

(1) 操作步骤

① 再次展开 Discriminant Analysis 对话框，选择分析变量和分类变量的操作同 1。

②在主对话框中，单击 Classify 按钮，展开 Classification 对话框选择分类参数：

- 在 Prior Probabilities 栏中选择 All groups equal 项。
- 在 Use Covariance Matrix 栏中选择 Within-groups 项。
- 在 Plots 栏中选择输出的统计图。选择 Combined-groups，要求作综合散点图。选择 Separate -groups，要对每类作一个散点图。选择 Territorial map，要求作区域图。
- 在 Display 栏中选择 Summary table，要求输出有关分类的数据。

③ 在主对话框中，单击 Statistics 按钮，展开相应对话框，选择要求输出的统计量。

• 在 Descriptives 栏中选择要输出的统计量：选择 Means 项，要求输出均值、标准差。选择 Univariate ANOVAs 要求输出每个变量的方差分析结果。检验的假设是：各类中同一自变量均值都相等。

- 在 Function Coefficients 栏中选择判别函数系数：选择 Fisher 要求输出费雪系数，

选择 **Unstandardized** 要求输出未标准化的判别函数的系数。

- 在 **Matrices** 栏中选择要输出的矩阵。选择 **Within-groups correlation** 要求显示合并类内相关矩阵，选择 **Within-groups covariance** 要求显示合并类内协方差矩阵，选择 **Separate-groups covariance** 要求显示各类的协方差矩阵，选择 **Total covariance** 要求显示总协方差矩阵。

④ 在主对话框中单击 **Save** 按钮，展开 **Save New Variables** 对话框。选择要求保存在工作数据文件中的新变量：选择 **Predicted group membership** 要求建立表明预测的类成员号的新变量，选择 **Discriminant Scores** 要求建立表明判别分数的新变量，选择 **Probabilities of group membership** 要求建立表明观测量作为各组成员的概率。

#### ⑤ 两点说明

- 由于在主对话框中仍然是选择了 **Enter independents together** 项，因此不能对判别分析方法进行进一步的选择。分析变量为所有 4 个反映鸢尾花特点的关于花瓣长、宽以及花萼长宽的变量。

- 使用所有观测量进行判别分析，同时因为工作数据文件中没有一个表示选择观测量的变量，因此无需单击 **Select** 按钮，对观测量进行进一步的选择。

(2) 主对话框中单击 **OK** 按钮，提交执行，执行的命令程序如下：

DISCRIMINANT	①
/GROUPS=spno(1 3)	②
/VARIABLES=slen swid plen pwid	③
/ANALYSIS ALL	④
/SAVE=CLASS SCORES PROBS	⑤
/PRIORS EQUAL	⑥
/STATISTICS=MEAN STDDEV UNIVF COEF RAW CORR COV GCOV TCOV TABLE	⑦
/PLOT=COMBINED SEPARATE MAP	⑧
/CLASSIFY=NONMISSING POOLED.	⑨

#### 命令语句说明

语句①②③④⑥⑨都是采用系统默认值时的语句。不再加以说明。

语句⑤是 **SAVE** 对话框中的选项生成的，要求生成新变量。选项 **CLASS** 要求建立一个表明各观测量属于哪一个类成员的变量。**SCORE** 要求建立三个变量，其值表明各观测量属于各类的类分数。**PROB** 要求建立三个变量表明各观测量属于各类的概率。

语句⑦是 **Statistics** 对话框中选项生成的。要求输出的原始数据的统计量有：**MEAN** 各变量均值和总均值，**STDDEV** 标准差（各变量的标准差和总标准差），**UNIVF** 单变量方差分析结果，**COEF** 判别函数的系数，**RAW** 未标准化的（原始）的判别函数系数，**CORR** 合并类间相关矩阵，**COV** 合并类间协方差阵，**GCOV** 各类协方差阵，**TCOV** 总协方差阵。

语句⑧是 **PLOT** 生成的。子命令要求输出三种统计图。**SEPARATE** 要求对每一类作



一个散点图，COMBINED 要求作三类合并在一张图上的散点图，MAP 要求作区域图。

语句@CLASSIFY=NONMISSING POOLED 子命令给出在预测分类时的选项。选项 NONMISSING 指定缺失值不参与分析，POOLED 要求使用合并类间协方差矩阵对工作数据文件中的观测量进行分类。

(3) 输出结果见表 13-30～表 13-37、图 13-41、图 13-42。结果与例 8 相同的表格不再重复列出。

分析的输出结果解释如下：

表 13-30 是原始数据描述统计量，包括基本数据信息外，还有各类中各变量的均值、标准差和整个样本的总均值、总标准差。

表 13-31 是各组均值相等的检验结果。进行的检验假设：各类中同变量均值相等。如果假设成立，说明根据各判别变量所作的原始分类是没有实际意义的。要么是分类错误，要么是选作判别的自变量不能充分显示分类特征。无论什么原因，进一步的输出结果分析均是无意义的。

如果有的变量方差分析结果表明变量对判别分析有意义，有的变量对判别分析无意义，则要改变判别分析方法，以便自动剔除对判别分析无意义的变量。

如果拒绝假设，说明原始分类有意义。同时，可以认为判别自变量能够表明分类特征。本例的方差分析结果 Sig.值均小于 0.001，说明 4 个判别变量都能很好地体现分类特征。但这并不说明所有变量相互独立，都应该出现在判别函数中。

表11-31 各组均值相等的检验

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
花萼长	.397	111.847	2	147	.000
花萼宽	.598	49.371	2	147	.000
花瓣长	.059	1.179E3	2	147	.000
花瓣宽	.071	960.007	2	147	.000

表 11-30 原始数据的描述统计量

Group Statistics					
分类		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
刚毛鸢尾花	花萼长	50.06	3.525	50	50,000
	花萼宽	34.28	3.791	50	50,000
	花瓣长	14.62	1.737	50	50,000
	花瓣宽	2.46	1.054	50	50,000
变色鸢尾花	花萼长	59.36	5.162	50	50,000
	花萼宽	27.66	3.147	50	50,000
	花瓣长	42.60	4.699	50	50,000
	花瓣宽	13.26	1.978	50	50,000
弗吉尼亚鸢尾花	花萼长	66.38	7.128	50	50,000
	花萼宽	29.82	3.218	50	50,000
	花瓣长	55.60	5.540	50	50,000
	花瓣宽	20.26	2.747	50	50,000
Total	花萼长	58.60	8.633	150	150,000
	花萼宽	30.59	4.363	150	150,000
	花瓣长	37.61	17.682	150	150,000
	花瓣宽	11.99	7.622	150	150,000

表 11-32 合并类内相关阵和协方差阵

Pooled Within-Groups Matrices <sup>a</sup>					
Covariance	花萼长	花萼宽	花瓣长	花瓣宽	
	花萼长	29.960	8.767	16.129	4.340
	花萼宽	8.767	11.542	5.033	3.145
	花瓣长	16.129	5.033	18.597	4.287
Correlation	花萼长	花萼宽	花瓣长	花瓣宽	
	花萼长	1.000	.471	.683	.387
	花萼宽	.471	1.000	.344	.452
	花瓣长	.683	.344	1.000	.486
	花瓣宽	.387	.452	.486	1.000

a. The covariance matrix has 147 degrees of freedom.

表 13-32 是合并类内相关矩阵（Correlation）和协方差矩阵（Covariance）。合并类内协方差阵各元素的值是各类协方差阵相应元素值之平均值。合并类内相关阵各元素的值是各类相关阵相应元素值之平均值。由此可以看出花瓣长和花萼长之间协方差值（16.129）和相关系数值（0.683）比较大，这就可以提出一个问题：是否它们之间不独立，在求出的判别函数中可否剔除一个变量呢？

表 13-33 是各类协方差阵和总协方差阵。除刚毛鸢尾花外，其余两种鸢尾花的协方

差矩阵中协方差系数（除自协方差外）最大的是花瓣长和花萼长之间的协方差值，为 18.290 和 28.461，Total 栏中的结果也一样为 130.036，因此有进一步分析的必要。

表 13-33 各类协方差阵和总协方差阵

Covariance Matrices <sup>a</sup>					
分类		花萼长	花萼宽	花瓣长	花瓣宽
刚毛鸢尾花	花萼长	12.425	9.922	1.636	1.033
	花萼宽	9.922	14.369	1.170	.930
	花瓣长	1.636	1.170	3.016	.607
	花瓣宽	1.033	.930	.607	1.111
变色鸢尾花	花萼长	26.643	8.288	18.290	5.578
	花萼宽	8.288	9.902	8.127	4.049
	花瓣长	18.290	8.127	22.082	7.310
	花瓣宽	5.578	4.049	7.310	3.911
弗吉尼亚鸢尾花	花萼长	50.812	8.090	28.461	6.409
	花萼宽	8.090	10.355	5.804	4.456
	花瓣长	28.461	5.804	30.694	4.943
	花瓣宽	6.409	4.456	4.943	7.543
Total	花萼长	74.537	-4.683	130.036	53.507
	花萼宽	-4.683	19.036	-33.056	-12.083
	花瓣长	130.036	-33.056	312.670	129.803
	花瓣宽	53.507	-12.083	129.803	58.101

a. The total covariance matrix has 149 degrees of freedom.

表 11-34 典则判别函数的系数

Canonical Discriminant Function Coefficients		
	Function	
	1	2
花萼长	-.063	.007
花萼宽	-.155	.218
花瓣长	.196	-.089
花瓣宽	.299	.271
(Constant)	-2.526	-6.987

Unstandardized coefficients

表 13-34 给出未标准化的典则判别函数的系数由于是未标准化的典则判别函数，因此有常数项，从表中可以得出两个判别函数分别是：

$$y_1 = -0.063 \times \text{花萼长} - 0.155 \times \text{花萼宽} + 0.196 \times \text{花瓣长} + 0.299 \times \text{花瓣宽} - 2.526$$

$$y_2 = 0.007 \times \text{花萼长} + 0.218 \times \text{花萼宽} - 0.089 \times \text{花瓣长} + 0.271 \times \text{花瓣宽} - 6.987$$

根据这两个典则判别函数可以计算出判别分数，根据各观测量的两个判别分数可以画出区域图或散点图。

有关判别函数的信息共输出 4 个表，典则判别函数系数表，类中心的函数值表，表明自变量与函数之间相关的结构矩阵表和非标准化的典则判别函数系数表。前三个表分别与表 13-31、表 13-32 和表 13-33 相同，在此不再列出与说明。

表 13-35 是分析中使用的各类的先验概率。由于在 Classification 对话框中现在的是各组先验概率相等，因此各为 0.333，分析中使用的观测量数加权、未加权都是 50。

表 13-36 用判别函数对观测量分类的结果。显示了费雪线性判别函数的系数。根据系数表可以总结出各类判别函数如下：

$$\text{刚毛鸢尾花: } F_1 = 1.687 \times \text{花萼长} + 2.695 \times \text{花萼宽} - 0.880 \times \text{花瓣长} - 2.284 \times \text{花瓣宽} - 80.268$$

$$\text{变色鸢尾花: } F_2 = 1.101 \times \text{花萼长} + 1.070 \times \text{花萼宽} + 1.001 \times \text{花瓣长} + 0.197 \times \text{花瓣宽} - 71.196$$

$$\text{弗吉尼亚鸢尾花: } F_3 = 0.865 \times \text{花萼长} + 0.747 \times \text{花萼宽} + 1.647 \times \text{花瓣长} + 1.695 \times \text{花瓣宽} - 103.896$$

使用 Fisher 判别函数的方法是测得一种鸢尾花的 4 个自变量：花萼长、花萼宽、花瓣长、花瓣宽的值，将 4 个自变量值代入上述 3 个函数式，得到 3 个函数值。比较这 3 个函数值，哪个值大就可以判断被测量的花属于哪类鸢尾花。例如：第一个观测量：花萼长 slen=50，花萼宽 swid=33，花瓣长 plen=14，花瓣宽 pwid=2。

代入函数 1，得到  $F_1=75.751$ 。  
代入函数 2，得到  $F_2=33.572$ 。  
代入函数 3，得到  $F_3=-9.547$ 。

比较 3 个值，可以看出  $F_1=75.751$  最大，据此得出第一个观测量属于刚毛鸢尾花。

表 13-37 是预测分类的小结。是一个判别回代小结。所谓回代就是对一个被测试的观测量使用下述方法判别属于的类：

- 使用除该观测量以外的观测量，求出线性判别函数；
- 使用求出的判别函数对这个观测量进行判别得出该观测量属于哪一类；
- 对每个观测量均使用该方法进行判别，然后统计错判率。与原始数据中的 spno 变量值进行比较得出错判概率。

表 13-35 分析中的先验概率

Prior Probabilities for Groups			
分类	Prior	Cases Used in Analysis	
		Unweighted	Weighted
刚毛鸢尾花	.333	50	50.000
变色鸢尾花	.333	50	50.000
弗吉尼亚鸢尾花	.333	50	50.000
Total	1.000	150	150.000

表 13-36 各类的分类函数的系数

	Classification Function Coefficients		
	分类		
	刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花
花萼长	1.687	1.101	.865
花萼宽	2.695	1.070	.747
花瓣长	-.880	1.001	1.647
花瓣宽	-2.284	.197	1.695
(Constant)	-80.268	-71.196	-103.890

Fisher's linear discriminant functions

表 13-37 预测分类结果小结

Classification Results <sup>a</sup>						
Original	Count	分类	Predicted Group Membership			Total
			刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花	
		刚毛鸢尾花	50	0	0	50
		变色鸢尾花	0	48	2	50
		弗吉尼亚鸢尾花	0	1	49	50
%		刚毛鸢尾花	100.0	.0	.0	100.0
		变色鸢尾花	.0	96.0	4.0	100.0
		弗吉尼亚鸢尾花	.0	2.0	98.0	100.0

a. 98.0% of original grouped cases correctly classified.

从表中可以看出利用判别函数回代的结果，刚毛鸢尾花的错判率为 0%；变色鸢尾花的错判率为 4%，是错判为第三类弗吉尼亚鸢尾花了，弗吉尼亚鸢尾花的错判率为 2%，有一个观测量被错判为第二类变色鸢尾花了。回代结果有 98% 判别正确。

图 13-41 是区域图。横坐标用第一个典则变量（Canonical Discriminant Function 1），纵坐标用第二个典则变量（Canonical Discriminant Function 2）。三种鸢尾花的典则变量值把一个典则变量组成的坐标平面划分成三个区域。可以看出变色鸢尾花的数据居于另外两种鸢尾花数据之间。

图 13-41 下面的列表是区域图中的标记符号。分别用 1、2、3 表明刚毛鸢尾花、变色鸢尾花、弗吉尼亚鸢尾花的区域，用“\*”表明各类鸢尾花的数据重心。

中心坐标用（典则判别函数 1 值，典则判别函数 2 值）表示这三种鸢尾花的中心分别为：刚毛鸢尾花中心（-7.392,0.219），变色鸢尾花中心（1.763, -0.737），弗吉尼亚鸢

尾花的中心 (5.629,0.518)。中心数据见表 13-34。

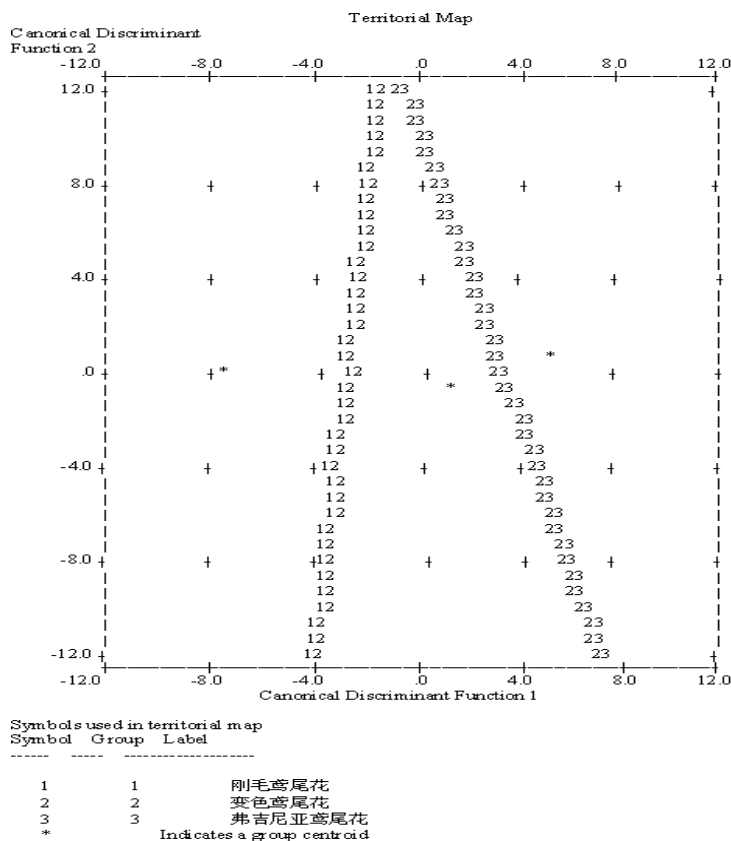
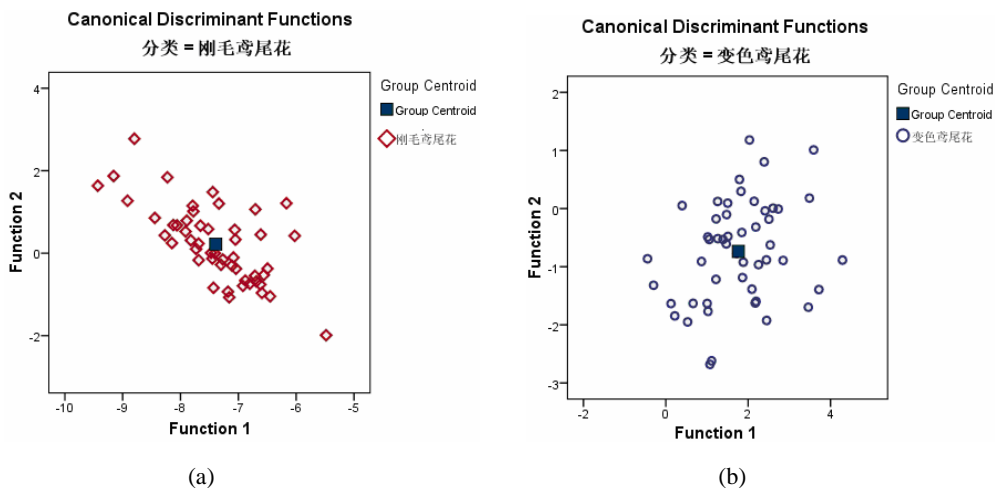


图 13-41 各类区域图及其标记说明



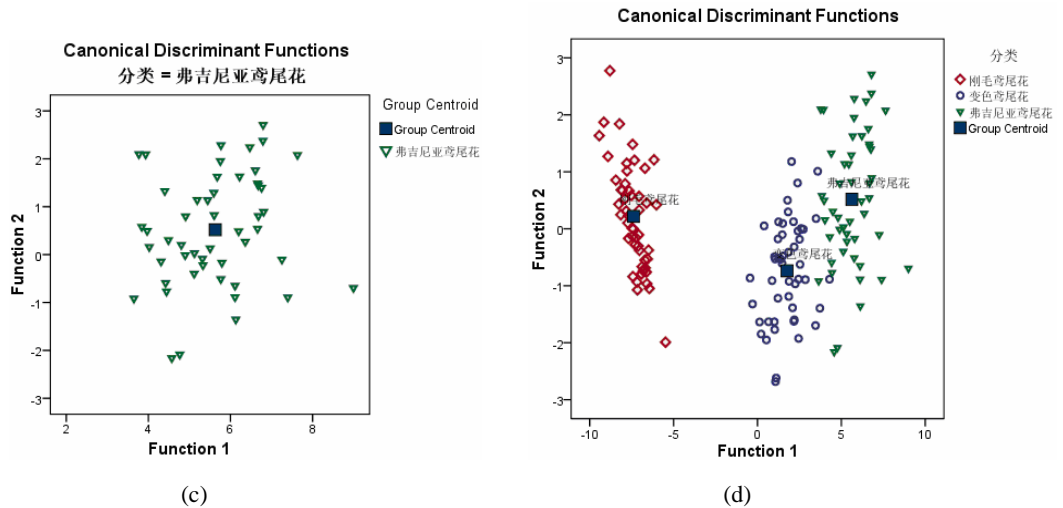


图 13-42 以典则判别函数为坐标的散点图

图 13-42 每个鸢尾花一个散点图，还有一个总的分类散点图。横坐标是典则判别函数 1，纵坐标是典则判别函数 2。根据自变量值计算两个典则判别函数值作出。总图中可以看出各类之间的关系。

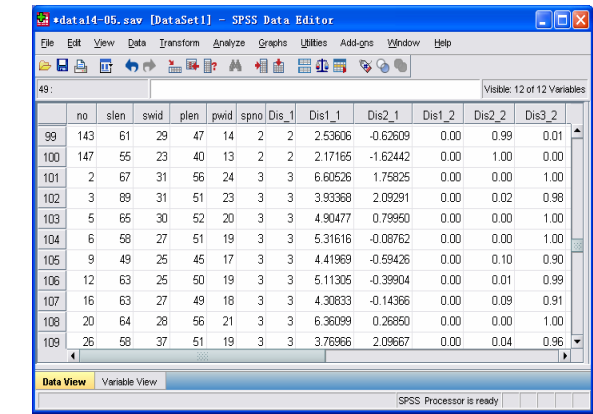


图 13-43 由 SAVE 子命令建立的新变量数据文件

在分析 1 中计算的未加权的典则变量 2 的值。

- 变量 DIS1\_2 表明各观测量在分析 1 中计算的属于第一类的概率。
- 变量 DIS2\_2 表明各观测量在分析 1 中计算的属于第二类的概率。
- 变量 DIS3\_2 表明各观测量在分析 1 中计算的属于第三类的概率。

如果分析方法不变，而且第一次运行产生的新变量没有从工作数据文件中删除，那

用 SAVE 子命令建立的新变量的信息表显示在输出文本的最开始处。新变量在数据窗中的情况见图 13-43。

新变量名采用系统默认方法。第一次运行命令程序建立的新变量的变量名及其数值含义说明如下：

- 变量 DIS\_1 表明分析 1 预测的各观测量所属类。
- 变量 DIS1\_1 表明各观测量在分析 1 中计算的未加权的典则变量 1 的值。
- 变量 DIS2\_1 表明各观测量

么可以再运行一次 Discriminant 过程, 得到第二次分析生成的新变量。读者可以对比两次运行在工作数据文件中所列出的变量名, 可以总结出系统默认变量名的规律。

### 13.5.4 逐步判别分析与实例

#### 1. 关于逐步判别分析

当研究某一事物分类时, 往往对于哪些变量能够反映研究范围内事物的特性这一问题的认识还不够深刻, 因此所选择的进行判别分析的变量不一定都能很好地反映类间差异。逐步判别分析假设已知的各类均属于多元正态分布, 用逐步选择法选择最能反映类间差异的变量子集建立较好的判别函数。一个变量能否被选择为变量子集的成员进入模型主要取决于协方差分析的 F 检验的显著性水平。

逐步判别分析从模型中没有变量开始。每一步都对模型进行检测。把模型外的对模型的判别力贡献最大的变量加入到模型中, 同时考虑已经在模型中, 但又不符合留在模型中的条件的变量从模型中剔除。直到模型中所有变量都符合留在模型中的判据; 模型外的变量都不符合进入模型的判据时为止。

实际工作中应该把使用逐步判别分析选择变量的结果与在实践中对变量的认识相结合, 会得到很好的判别分析模型。

#### 2. 逐步判别分析方法与判据的选择

逐步判别操作步骤见本章 13.5.2 节的内容。在 Discriminant 主对话框中应该选择 Use stepwise method 项。单击该选项, 并进一步选择分析方法或判据。

单击 Method 按钮, 展开 Stepwise Method 逐步判别方法对话框, 如图 13-37 所示。在对话框中显示出系统默认的逐步判别方法是 Wilks'Lambda。其判据是: 进入模型的  $F$  值为  $F \geq 3.84$ ; 从模型中剔除变量的判据是  $F$  值为  $F \leq 2.71$ 。不熟悉统计分析的读者可以不再进一步选择, 直接使用系统默认的分析方法和判据。逐步判别方法和判据的选择以及要显示的输出内容均参见 13.5.2 节。

**【例 10】**为了容易比较, 仍用鸢尾花的数据 (data13-05) 作为逐步判别分析的数据。

例 8、例 9 中的程序都是使用全部变量建立判别函数。能否减少变量仍然得到较好的判别函数呢? 我们采用 Wilks'Lambda 方法进行逐步判别分析。使用  $F$  值作为判据统计量。当  $F \geq 30$  时变量进入模型; 当  $F \leq 5$  时, 变量从模型中移出。

##### (1) 操作步骤如下:

① 再次展开 Discriminant Analysis 对话框。

② 仍把全部 4 个自变量送入 Independents 框中, 变量 spno, 作为分类变量移到 Grouping Variable 框中。单击 Define Range 按钮, 在对话框中, 输入变量 spno 的数值范围, 最小值 1 和最大值 3。

③ 在主对话框中, 选择 Use stepwise method 项, 单击 Method 按钮, 展开 Stepwise method 对话框, 在 Method (方法) 栏中选择 Wilks'Lambda 项; 在 Criteria 栏中选择 Use

F value 项,并在 Entry 框输入 30, Remove 框中输入 5;在 Display(显示)栏中选择 Summary of step 要求显示逐步选择变量子集的小结。F for pairwise distance 要求显示每两类之间的成对的 F 矩阵。

④ 在主对话框中单击 Statistics 按钮,在统计量对话框中选择 Mean、Univariate ANOVA 项;选择 Fisher、Unstandardized、Within-groups correlations。

⑤ 在主对话框中单击 Classify 按钮,展开 Classification 对话框。选择 All groups equal 各组先验概率相等;选择 Within-groups 项,使用组内协方差矩阵;选择 Summary table,要求显示聚类回代结果的小结表。

⑥ 在主对话框中单击 Save 按钮,展开 Save New Variables 对话框。选择 Predicted group membership 生成预测的类别的新变量、Discriminant scores 生成判别函数的分数新变量和 Probabilities of group membership 生成观测量属于各类的概率的新变量。

⑦ 在主对话框中单击 OK 按钮,提交运行。

(2) 运算程序与例 8、例 9 中的程序基本相同,仅增加以下语句:

/METHOD=WILKS	①
/FIN= 30	②
/FOUT= 5	③
/HISTORY	④

(3) 程序语句解释

① 指定逐步判别分析使用 Wilks'Lambda 方法的 Method 子命令。

②、③两个子命令指定逐步判别分析使用  $F$  值作判据,进入模型的判别变量  $F$  值必须大于等于 30,从模型中剔除出的变量  $F$  值必须小于等于 5。

④ HISTORY 子命令要求显示判别分析对变量的每一步选择后,各变量统计量和选定变量子集后的变量选择小结。

(4) 输出结果见表 13-38~表 13-50。与例 8 重复的不再列出。

(5) 输出结果解释

① 原始变量的描述统计量表与表 13-34 相同。是各类自变量的均值与标准差与自变量的总的均值和标准差。

从各类均值与标准差的比较中可以看出各类鸢尾花中,变量花萼宽 swid 标准差值比其他变量值集中,总标准差最小。由此看到一个可能性:如果能从判别函数中减掉一个变量,这个变量可能是花萼宽 swid。但与花瓣宽、花萼长的标准差在一个数量级上。因此还是要经过逐步判别分析才能最后确定。

② 表 13-38 逐步判别分析前相关阵。从前面表 13-35 虽然可以看出四个自变量都对区分鸢尾花的种类是有效变量,但根据表 13-42 相关矩阵可以看出花瓣长和花萼长相关系数比较大,为 0.683。能否在判别函数中省掉一个自变量呢?这个问题由逐步判别分析来解决。

表 13-38 逐步判别前的自变量相关阵

Pooled Within-Groups Matrices					
	花萼长	花萼宽	花瓣长	花瓣宽	
Correlation	1.000	.471	.683	.387	
花萼宽	.471	1.000	.344	.452	
花瓣长	.683	.344	1.000	.486	
花瓣宽	.387	.452	.486	1.000	

表 13-39 逐步判别的分析小结

Variables Entered/Removed <sup>a,b,c,d</sup>									
Step	Entered	Wilks' Lambda				Exact F			
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	花萼长	.059	1	2	147.000	1179.052	2	147.000	.000
2	花萼宽	.038	2	2	147.000	301.876	4	292.000	.000
3	花瓣宽	.026	3	2	147.000	251.164	6	290.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 8.

b. Minimum partial F to enter is 30.

c. Maximum partial F to remove is 5.

d. F level, tolerance, or VIN insufficient for further computation.

表 13-40 逐步进入模型的变量方差分析结果

Variables in the Analysis				
Step		Tolerance	F to Remove	Wilks' Lambda
1	花萼长	1.000	1179.052	
2	花萼宽	.882	1078.565	.598
	花瓣长	.882	39.965	.059
3	花瓣宽	.745	36.018	.039
	花萼宽	.775	49.885	.044
	花瓣宽	.672	33.060	.038

表 13-41 各步模型外的变量方差分析结果

Variables Not in the Analysis				
Step		Tolerance	Min. Tolerance	F to Enter
0	花萼长	1.000	1.000	111.847
	花萼宽	1.000	1.000	49.371
	花瓣长	1.000	1.000	1179.052
	花瓣宽	1.000	1.000	960.007
1	花萼长	.533	.533	21.768
	花萼宽	.882	.882	39.965
	花瓣长	.764	.764	24.435
2	花萼长	.470	.470	6.733
	花瓣长	.672	.672	33.060
3	花萼长	.469	.469	4.159

③ 表 13-39 是逐步判别分析的一个小结。Exact F 栏内的 Statistic 是一个  $F$  值，是该变量的均方与误差均方的比值。该值越大，Sig 值越小，因此该值最大的先进入判别函数。当 Sig 小于 0.05 或 0.01 时，拒绝零假设。显著性检验结果 Sig=0.000，即小于 0.001，可以说这三个变量对判别的贡献都很显著。总之，说明该变量在不同类中均值不同是由于类间差异，而不是由随机误差引起的。既该变量在各组中均值差异显著。可以看出三个变量的  $F$  统计量都在 30 以上，这是我们选择的进入判别函数的判据。

表下面的 5 个注解是：

- 选择原则：使 Wilks' Lambda 最小的，即使该值降低最多的变量。
- 最大步数：8，结束迭代。
- 进入模型（变量子集）的最小  $F$  值：30。
- 移出模型的最大  $F$  值：5。
- $F$  检验的结果不能满足进一步的判别函数的计算，因此结束运算。这点可以从后面的输出表中的数据分析得出。

④ 表 13-40、表 13-41 是根据 Wilks' Lambda 值进行逐步选择变量并进行  $F$  检验的过程数据。每一步都计算该变量进入模型使 Wilks' Lambda 值降低了多少。都是那个使总的 Wilks' Lambda 值最小的变量进入判别函数。从这两个表可以看到逐步判别的每一步过程。判别分析在一个自变量进入模型后，对模型内各变量进行方差分析，在模型外的自变量进行方差分析和  $F$  检验。模型内的  $F$  检验  $F$  值小于 5 的自变量还要从模型中移出。模型外的自变量若  $F$  值大于 30，可以进入模型。



- 表 13-41 中 Step 0 中表明花瓣长  $F$  值最大,  $F=1179.052$ , Willk's Lambda=0.059 值最小, 第一个进入模型的是花瓣长。
- 进入模型后因为只有一个变量, 在表 13-40 中 Step=1 行中可以看出花瓣长第一个进入模型。

表 13-42 每步的类间比较

Pairwise Group Comparisons <sup>a,b,c</sup>					
Step	分类		刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花
1	刚毛鸢尾花	F		1052.420	2257.552
		Sig.		.000	.000
	变色鸢尾花	F	1052.420		227.185
		Sig.	.000		.000
	弗吉尼亚鸢尾花	F	2257.552	227.185	
		Sig.	.000	.000	
2	刚毛鸢尾花	F		768.305	1416.055
		Sig.		.000	.000
	变色鸢尾花	F	768.305		115.071
		Sig.	.000		.000
	弗吉尼亚鸢尾花	F	1416.055	115.071	
		Sig.	.000	.000	
3	刚毛鸢尾花	F		656.739	1316.404
		Sig.		.000	.000
	变色鸢尾花	F	656.739		129.425
		Sig.	.000		.000
	弗吉尼亚鸢尾花	F	1316.404	129.425	
		Sig.	.000	.000	

a. 1, 147 degrees of freedom for step 1.  
b. 2, 146 degrees of freedom for step 2.  
c. 3, 145 degrees of freedom for step 3.

表 13-43 典则判别函数的特征值表

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	28.708 <sup>a</sup>	99.0	99.0	.983
2	.292 <sup>a</sup>	1.0	100.0	.476

a. First 2 canonical discriminant functions were used in the analysis.

表 13-44  $\lambda$  值的卡方转换及卡方检验

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.026	532.603	6	.000
2	.774	37.454	2	.000

表 11-45 标准化的典则判别函数系数表

Standardized Canonical Discriminant Function Coefficients		
	Function	
	1	2
花萼宽	-.640	.758
花瓣长	.656	-.367
花瓣宽	.642	.549

表 13-46 结构矩阵

	Function	
	1	2
花瓣长	.747*	.160
花萼长 <sup>a</sup>	.395*	.319
花萼宽	-.125	.880*
花瓣宽	.671	.714*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions  
Variables ordered by absolute size of correlation within function.

\*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

表 13-47 未标准化的典则判别函数系数表

Canonical Discriminant Function Coefficients

	Function	
	1	2
花萼宽	-.188	.223
花瓣长	.152	-.085
花瓣宽	.314	.268
(Constant)	-3.715	-6.842

Unstandardized coefficients

表 13-48 各类中心的未标准化的判别函数值表

Functions at Group Centroids

分类	Function	
	1	2
刚毛鸢尾花	-7.180	.219
变色鸢尾花	1.708	-.737
弗吉尼亚鸢尾花	5.472	.518

Unstandardized canonical discriminant functions evaluated at group means

• 表 13-41 的第一步 Step 1, 花瓣长进入模型后, 模型外的三个自变量的方差分析, 花萼宽的  $F$  值最大, 其  $F$  值为 39.965, 大于 30, Willk's Lambda=0.038 值最小, 因此第二个进入模型的是花萼宽。

• 表 13-40 第二步 Step 2 花萼宽进入模型, 模型内方差分析结果: 花瓣长  $F=1078.565$ ; 花萼宽  $F=39.965$ , 都大于 5, 因此两个变量都保持在模型中。

• 表 13-41 第二步 Step 2 模型外的变量方差分析结果,  $F$  值最大的是花瓣宽,  $F=33.060$ , 也大于 30. Willk's Lambda=0.026 值最小, 应该自变量花瓣宽进入模型。

• 同样表 13-40 中的 Step 3 花瓣宽进入模型后的方差分析结果,  $F$  值均大于 5, 三个自变量仍保持在模型中。

• 表 13-41 中的 Step 3 模型外自变量方差分析, 花萼长  $F=4.159$ , 小于 30, 该自变量不再进入判别函数模型。

模型外、内变量无进、无出, 逐步判别分析的自变量选择结束。

⑤ 表 13-42 是在逐步判别分析过程的每一步中, 在任意两类之间进行的方差分析, 想看看在这一步选入模型中的自变量对任意两类之间的区分是否有效。 $F$  值越大, Sig 值越小, 区分效果越好。行类与列类之间的方差分析结果显示在行列交叉单元格中。可以从表中看出各步所选择的变量对任意两类的区分都是有效的。

输出窗口中在 Summary of Canonical Discriminant 标题下的表格说明了使用选择的自变量导出的典则判别函数的结果:

⑥ 表 13-43 为两个典则判别函数的特征值表。可以看出与全模型的特征值相差不多。第一个函数仍占了总方差的 99%。

⑦ 表 13-44 为 Wilks' Lambda 值的卡方转换及检验。

⑧ 表 13-45 为标准化的典则判别函数系数表。可从中总结出标准化的典则判别函数

$$y_1 = -0.640 \times \text{花萼宽} + 0.656 \times \text{花瓣长} + 0.642 \times \text{花瓣宽}$$

$$y_2 = 0.758 \times \text{花萼宽} - 0.367 \times \text{花瓣长} + 0.549 \times \text{花瓣宽}$$

注意, 若用标准化的判别函数计算标准化的判别分数, 必须代入标准化的自变量值。

⑨ 表 13-46 是判别自变量与标准化的典则判别函数之间的相关矩阵。标字母 a 的自变量没有在判别函数中。

⑩ 表 13-47 未标准化的典则判别函数:

$$y_1 = -0.188 \times \text{花萼宽} + 0.152 \times \text{花瓣长} + 0.314 \times \text{花瓣宽} - 3.715$$

$y_2=0.223\times\text{花萼宽}-0.085\times\text{花瓣长}+0.268\times\text{花瓣宽}-6.842$

使用未标准化的典则判别函数计算判别分数，使用原始自变量值。

表 13-49 逐步判别选择的变量进行线性判别分析结果

Classification Function Coefficients			
	分类		
	刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花
花萼宽	3.452	1.564	1.135
花瓣长	.411	1.843	2.309
花瓣宽	-2.425	.105	1.622
(Constant)	-60.280	-62.685	-98.632

Fisher's linear discriminant functions

表 13-50 逐步判别回代小结

Classification Results <sup>a</sup>					
		Predicted Group Membership			Total
		刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花	
Original	Count				
	刚毛鸢尾花	50	0	0	50
	变色鸢尾花	0	48	2	50
%	刚毛鸢尾花	100.0	.0	.0	100.0
	变色鸢尾花	.0	96.0	4.0	100.0
	弗吉尼亚鸢尾花	.0	2.0	98.0	100.0

a. 98.0% of original grouped cases correctly classified.

⑪ 表 13-48 各类中心处的未标准化的典则判别函数值。与区域图中的“\*”坐标值对应。在区域图中以两个典则判别函数值为两个坐标轴，平面上的点表示为 $(f_1,f_2)$ 则各种鸢尾花的类中心坐标分别为：刚毛鸢尾花： $(-7.180, 0.219)$ ；变色鸢尾花： $(1.708, -0.737)$ ；弗吉尼亚鸢尾花： $(5.472, 0.518)$ 。

输出窗口中，在 Classification 标题下的表格是使用判别函数对原始数据进行分类的结果数据。

⑫ 表 13-49 是使用逐步判别选择的变量进行线性判别分析结果。逐步判别选择变量的目的仍是要使用选择出的较少的自变量，导出判别函数，对观测量进行进一步的判别，然后分析该判别函数的优劣。表 13-53 是 Fisher 线性判别函数系数表。三个线形判别函数如下：

$F_1=3.452\times\text{花萼宽}+0.411\times\text{花瓣长}-2.425\times\text{花瓣宽}-60.280$   
 $F_2=1.564\times\text{花萼宽}+1.843\times\text{花瓣长}+0.105\times\text{花瓣宽}-62.685$   
 $F_3=1.135\times\text{花萼宽}+2.309\times\text{花瓣长}+1.622\times\text{花瓣宽}-98.632$

使用线性判别函数，应该代入各判别变量原始观测值，计算判别函数值即判别分数。

⑬ 表 13-50 逐步判别回代小结。

从表中数据可以看出，该表是用只包含三个变量的判别函数进行分类的小结。可以看出对刚毛鸢尾花的分类错判率为 0%；对变色鸢尾花的分类有 2 个观测量错判为弗吉尼亚鸢尾花了，错判率为 4%。对弗吉尼亚鸢尾花错判了 1 个，错判率为 2%。总的判断正确率为 98%（错判率为 2%）。虽然比起全模型来少了一个自变量，但错判率没有改变。

由此也说明逐步判别的结果可行。

读者可以在 Plot 选项中选择分类散点图、各类散点总图和区域图对各结果进行进一步的认识。由于篇幅关系,不再一一列出。

### 13.5.5 判别分析过程的命令语句

判别分析过程 Discriminant 过程可以使用下列命令语句。

#### 1. 命令语句格式:

```
DISCRIMINANT GROUPS=varname(min,max) /VARIABLES=varlist
[/SELECT=varname(value)] [/ANALYSIS=varlist[(level)] [varlist...]]
[/METHOD={DIRECT**}{WILKS}{MAHAL}{MAXMINF}{MINRESID}{RAO}]
[/TOLERANCE={0.001}{n}] [/MAXSTEPS={n}] [/FIN={3.84**}{n}]
[/FOUT={2.71**}{n}] [/PIN={n}] [/POUT={n}] [/VIN={0**}{n}]
[/FUNCTIONS={g-1,100.0,1.0**}{n 1,n 2,n 3}]
[/PRIORS={EQUAL**}{SIZE}{value list}]
[/SAVE=[CLASS=[varname]] [PROBS=[rootname]] [SCORES=[rootname]]]
[/ANALYSIS=...] [/MISSING={EXCLUDE**}{INCLUDE}]
[/MATRIX=[OUT({*}){file}] [IN({*}){file}]] [/HISTORY={STEP**}{NONE}]
[/ROTATE={NONE**}{COEFF}{STRUCTURE}]
[/CLASSIFY={NONMISSING}{POOLED}{UNSELECTED}{SEPARATE} [UNCLASSIFIED]
[MEANSUB]]
[/STATISTICS=[MEAN] [STDDEV] [CORR] [COV] [GCOV] [TCOV] [COEFF]
[UNIVF] [RAW] [FPAIR] [BOXM] [TABLE] [CROSSVALID] [ALL]]
[/PLOT=[MAP] [SEPARATE] [COMBINED] [CASES(n)] [ALL]]
```

其中: **DISCRIMINANT** 是命令关键字, **GROUP**=后面的变量是已知的分类变量。变量名后面的括号中的两个值说明各变量值的范围, 前面一个是最小值, 后面一个是最大值, 中间用逗号分隔开。调用 **DISCRIMINANT** 过程必须用 “**GROUPS=**” 指定一个分类变量。

“[/]” 中的均为子命令, 一个子命令中的若干个并列的 “[ ]” 中的内容表明该子命令完成其功能需要指定的参数。是子命令下属的第一层选项, 并列的几个选项可以取其一, 也可以取其中的若干项; 在这一层选项下面并列的若干个用 “{ }” 包含的内容是第二层选项, 并列的选项中只能取其一。

#### 2. 子命令及其含义

(1) **VARIABLES** 子命令指定判别分析中使用的自变量。在等号后面列出自变量表, 各变量名之间用空格分隔开。如果不使用此子命令, 则使用数据文件中除 **GROUPS=** 指定的分类变量以外所有的数值型变量作为自变量。

(2) **SELECT** 子命令指定筛选参与分析的观测量。在等号后面写出筛选变量的变量名,并在其后的括号中写出筛选值。**DISCRIMINANT** 过程只使用该变量值为括号中指定值的观测量进行分析。不使用此子命令,数据文件中的全部观测量均参与判别分析。

(3) **ANALYSIS** 子命令指定特殊分析使用的变量。这些变量必须已经在 **VARIABLES** 子命令中指定过,变量名后面的括号中给出该变量的容差水平。一般情况下使用所有 **VARIABLES** 子命令指定的变量进行判别分析。该子命令写成:“/ANALYSIS=ALL”。

(4) **METHOD** 子命令。不使用该子命令,则使用 **VARIABLES** 子命令指定的所有变量推导全变量的判别函数。使用 **METHOD** 子命令,则使用选择的方法进行判别分析,选择最能反映各类观测量之间差别的变量子集,利用子集中的变量进行判别分析。可以选择的判别分析方法有以下六种,无论指定哪种方法均应考虑指定选择变量工作停止的判据。指定判据也有相应的子命令。

- **WILKS**, 指定用 Wilks' Lambda 方法, Wilks'  $\lambda$  值最小的变量进入模型。
- **MAHAL**, 指定使用 Mahalanobis distance 方法,使最近的两类之间的 Mahalanobis 距离最大的变量进入模型。
- **MAXMINF**, 指定使用 Smallest F ratio 法,即使任意两类间的最小的  $F$  比值最大。
- **MINRESID**, 指定使用最小残差法。
- **RAO**, 指定使用 Rao's  $V$  法,即选入变量要求 Rao's  $V$  统计量最大。使用此种方法选择变量要同时使用 **VIN** 子命令指定增量  $V$ , 否则使用默认值  $V=0$ 。
- **DIRECT**, 指定使用直接法。是推导全变量模型的判别函数时,不指定 **METHOD** 子命令的默认方法。

(5) **TOLERANCE** 子命令。该子命令指定容差,默认值是 0.001。

(6) **MAXSTEP** 子命令。指定一个逐步选择变量的过程停止的判据。该判据用最大步数确定。默认值为自变量数目的二倍。也可以在等号后面指定该数值。

(7) **FIN** 子命令。当使用  $F$  统计量作判据时,用 **FIN** 子命令指定变量进入模型的最小  $F$  值。当模型外的变量大于 **FIN** 指定的  $F$  值时,变量可以进入模型。**FIN** 指定的值应该大于 **FOUT** 指定的  $F$  值。默认的 **FIN** 值为 3.84。格式中的  $n$  表示由读者指定的值。

(8) **FOUT** 子命令。当使用  $F$  统计量作判据时,指定变量移出模型的最大  $F$  值。当变量的  $F$  值小于 **FOUT** 子命令指定的  $F$  值时,变量从模型中移出。**FOUT** 指定的  $F$  值应该小于 **FIN** 指定的  $F$  值。否则,模型中将不会有变量存在。默认的 **FOUT** 值为 2.71。格式中的  $n$  表示由读者指定的值。

(9) **PIN** 子命令。当使用“大于  $F$  值的概率”作判据时,使用该子命令指定进入模型的  $F$  的概率。只有变量的  $F$  值的概率小于这个指定值时,变量才可进入模型。

(10) **POUT** 子命令。当使用“大于  $F$  值的概率”作判据时,使用该子命令指定从模型中剔除变量的  $F$  的概率。只有变量的  $F$  值概率大于这个指定值时,变量从模型中移出。

(11) **VIN** 子命令。当使用了 **METHOD=RAO** 时用 **VIN** 子命令指定变量进入模型的

最小增量。默认值为 0。

(12) **FUNCTIONS** 子命令。该子命令指定推导典则判别函数（典则变量）使用的参数，共 3 个，括号中的三个参数之间应该用逗号分隔开。这些参数依次为：

- **nf**，函数数目的最大值。默认值是分类变量水平数减 1 ( $g-1$ )。
- **cp**，方差累计百分比最小值。默认值是 100。
- **sig**，Wilks'  $\lambda$  的最大显著性水平。默认值是 1.0。

(13) **PRIORS** 子命令指定进行判别分析时使用的各类的先验概率。指定方法有 3 种：

- **EQUAL**，指定各类先验概率相等。每一类的先验概率为总类数的倒数。
- **SIZE**，指定先验概率与各类观测量数目成正比。即第  $i$  类的先验概率为  $n_i/\sum n_i$ 。
- 先验概率表给出各类先验概率值，顺序应该与 **GROUP** 子命令定义的变量值范围的顺序一致。各概率值之间应该用逗号分隔开。各类先验概率之和应该等于 1。

(14) **SAVE** 子命令建立保存在工作数据文件中的新变量。可以选择的变量有 3 种：

① **CLASS**=变量名，新变量的值为判别结果，即每个观测量所属的类号。变量名可以由读者指定，必须由小于或等于 8 个 **SPSS** 允许的字符组成。不指定变量名则该选项只写“**CLASS**”，系统给出默认的变量名形式为 *dis\_n*， $n$  为建立同一变量的各次运行顺序号。

② **SCORE**=变量名字头，新变量的值为表明各观测量属于各类的判别分数，是一组变量。等号后面只需给出新变量名的字头，由系统自动在指定字头后面加类别序号。指定的变量名字头加上类序号组成的变量名不能超出 7 个字符，否则在输出窗口中给出错误信息，子命令不能执行。该组变量的数目等于已知的分类的数目减 1。不指定变量名字头，系统给出默认的变量名形式为 *disn\_m*， $n$  为用第  $n$  个未加权的典则判别函数计算的判别分数。 $m$  为多次运行时建立该组变量时的运行顺序号。如果该选项与 **PROB** 选项同时选用， $m$  值为运行顺序号乘 2 减 1（奇数）。

③ **PROB**=变量名字头，新变量值表明经过判别函数判别，观测量属于各类的概率。该组变量的数目与分类数目相等。变量名字头的命名方法与 **SCORE** 子命令相同。系统自动给出的默认变量名形式为 *disn\_m*， $n$  为属于第  $n$  类的概率， $m$  为多次运行 **DISCRIMINAT** 过程均建立该组变量时的运行顺序号。如果该选项与 **SCORE** 选项同时选用， $m$  值为运行顺序号乘 2（偶数）。

注意：读者自己命名变量名或变量名字头时可以使用 26 个英文字母和“.”、“\_”。字符数应该小于等于 7。

(15) **MISSING** 子命令指定处理带有缺失值的观测量的方法。共有两种选择：

- **EXCLUDE**，指定进行判别分析时剔除带有缺失值的观测量。
- **INCLUDE**，指定进行判别分析时不剔除带有缺失值的观测量。

(16) **MATRIX** 子命令指定进行判别分析使用的相关矩阵文件和分析后输出的相关矩阵文件。只有当不使用原始数据时使用 **IN** 选项指定一个输入矩阵文件，或需要将中间

结果作他用时使用 OUT 选项。一般情况下很少使用该子命令。选项 IN 或 OUT 后面指定的文件名应该是带有路径的完整的文件名。文件名置于括号中。

如果当前的工作数据文件就是一个相关矩阵或由相关分析得出的相关矩阵,则在 IN 选项中使用“\*”作为默认文件名,分析时使用当前工作数据文件(数据编辑窗口中的矩阵)。如果将判别分析中的相关矩阵输出到数据编辑窗口,则将 OUT 选项中使用“\*”做默认的数据文件名。

(17) HISTORY 子命令指定是否输出判别分析的每一步和最后结果。选项有两组:

STEP、NON 二者选其一,前者为默认值。STEP 指定按步骤给出逐步选择变量的结果, NON 不显示中间过程。

(18) ROTATE 子命令指定是否对判别函数系数矩阵或结构矩阵进行旋转。一般使用默认方法即不进行旋转。

(19) CLASSIFY 子命令指定对观测量进行回代分类时的范围或方法:

- NOMISSING, 指定只对不带有缺失值的观测量进行回代分类。
- UNSELECTED, 指定对未被选入参与判别分析的观测量(即 SELECT 子命令未包括的观测量)进行分类。
- UNCLASSIFIED, 指定只对数据文件中包含的未分类的观测量进行分类。
- POOLED, 指定使用合并类内协方差矩阵。
- SEPARATE, 指定使用类间协方差矩阵。

(20) STATISTICS 子命令指定输出的统计量,常用的有:

- MEAN 均值, 包括各类的各变量均值和各变量的总均值。
- STDDEV 标准差。分类给出各变量的标准差, 还给出各变量的总的标准差。
- CORR, 类内相关矩阵。
- COV、GCOV、TCOV 三个选项依次为: 类内协方差矩阵、类间协方差矩阵、总协方差矩阵。

- COEFF, 线性判别函数系数。

- UNIVF, 包括各变量的 Wilks'λ 值、F 值和显著性概率。

- RAW, 未标准化的典则判别函数系数。

- FPAIR, 两两类之间的 F 值矩阵。

- BOXM, 针对各类协方差矩阵相等这一假设的 Box M 检验结果。

- TABLE, 要求给出回代的分类结果表。

还有一个特殊的选项即“ALL”, 表示要求以上所有选项指定的输出。

(21) PLOT 子命令指定要输出的统计图, 可以指定两种不同输出方式的统计图, 所有统计图均用典则判别函数即典则变量作为坐标轴。选项有 5 个:

- MAP, 区域图, 表明各类数值区域的图。

- SEPARATE, 散点图, 每一类用一张散点图表明观测量分布情况。观测量共分为

几类，就产生几张散点图。

- COMBINED，散点图，各类观测量合并在一个散点图中，显示其观测量分布，标明类中心。
- ALL，要求作出所有统计图。

## 习 题 13

1. SPSS 提供几种聚类分析过程？各适合什么情况的聚类？
2. 聚类分析与判别分析对数据要求有什么不同？
3. 聚类分析之前一定要对变量进行标准化吗？为什么？
4. 变量聚类后如何根据聚类结果确定各类的代表变量？
5. 1976 年 74 个国家人口出生率和死亡率数据在 data13-06.xls 中。将数据转换成 SPSS 数据文件，以相同的主名保存成.sav 文件，根据出生率、死亡率聚类，绘制散点图。
6. 数据 data13-07.xls 中 sheet1 中是 28 名一级，25 名健将级标枪运动员测验的 6 项影响标枪成绩的项目成绩。据此求出判别运动员等级的判别函数。回代，给出错判率。Sheet2 中是 14 名未知级别的运动员。运用判别函数对他们分类。转换成的 SPSS 数据文件请参考 data13-07.sav 和 data13-07b。
7. 上述 6 个与标枪成绩有关的项目彼此是否相关？能否进行变量聚类，并找出各类中有代表性的项目（变量）？
8. 用逐步判别法再求判别函数，与用全部变量求出的判别函数比较错判率。



## 第 14 章 因子分析与对应分析

在各个领域的科学研究中，往往需要对反映事物的多个变量进行大量的观测，收集大量数据以便进行分析寻找规律。多变量大样本无疑会为科学研究提供丰富的信息，但也在一定程度上增加了数据采集的工作量，更重要的是在大多数情况下，由于许多变量之间可能相关，增加了问题分析的复杂性，同时对分析带来不便。如果分别分析每个指

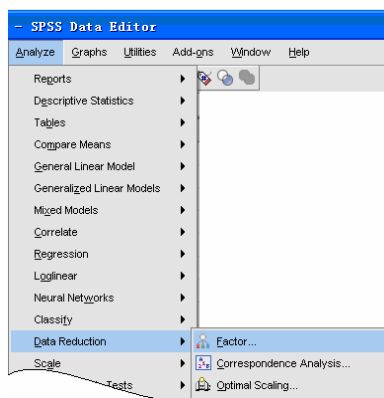


图 14-1 调用因子分析过程命令

标，分析又可能是孤立的，而不是综合的。盲目减少指标会损失很多信息，容易产生错误的结论。因此需要找到一个合理的方法，减少分析指标的同时，尽量减少原指标包含信息的损失，对所收集的资料作全面的分析。由于各变量间存在一定的相关关系，因此有可能用较少的综合指标，分别综合存在于各变量中的各类信息。这就是降维。SPSS 收集的降维方法在 Analyze 的 Data Reduction 菜单中。Factor（因子分析）、Correspondence Analysis（对应分析）和 Optimal Scaling（最优尺度）分析就是这样的降维方法，见图 14-1。

### 14.1 主成分分析与因子分析

#### 14.1.1 主成分分析与因子分析概述

##### 1. 主成分分析的概念

##### (1) 什么是主成分分析

在各领域的科学研究中，为了全面客观地分析问题，往往要考虑从多方面观察所研究的对象，要收集多个观察指标的数据。如果一个一个地分析这些指标，无疑会造成对研究对象的片面认识，也不容易得出综合的、一致性很好的结论。主成分分析就是考虑各指标间的相互关系，利用降维的思想把多个指标转换成较少的几个互不相关的综合指标，从而使进一步研究变得简单的一种统计方法。

现举例说明主成分分析。儿童身高和体重两个变量之间的关系（见表 14-1），可以使用散点图表示出来，如图 14-2 所示。显然，这两个变量之间存在线性关系。数据  $(h_i, w_i)$

各点散布在一条直线周围, 其中  $i=1\sim n$ 。

现在以该直线为一个坐标轴  $p_1$ , 以该轴的垂直线为另一个坐标轴  $p_2$ 。因为所有观测点均在坐标轴  $p_1$  周围, 而  $p_1$  与  $p_2$  是两个相互垂直的坐标轴, 因此彼此不相关。

原观测点可以表示为  $(p_{1i}, p_{2i})$ ,  $i=1\sim n$ 。可以认为,  $n$  个观测的差异主要表现在  $p_1$  方向上, 而在  $p_2$  方向上差异很小。

表 14-1 身高体重数据

变量 观测量 I	身高 h	体重 w
1	$h_1$	$w_1$
2	$h_2$	$w_2$
3	$h_3$	$w_3$
4	$h_4$	$w_4$
...	...	...
$n$	$h_n$	$w_n$

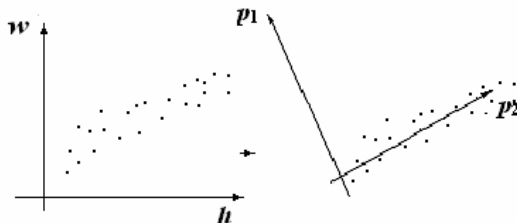


图 14-2 主成分概念示意图

由此可以得出结论, 可以用  $p_1$  一个指标来代替原始变量  $h$ 、 $w$  研究  $n$  个观测对象的差异。  $p_1$ 、 $p_2$  可以用原始变量  $h$ 、 $w$  的线性组合来表示

$$\begin{cases} p_1 = l_{11}h + l_{12}w \\ p_2 = l_{21}h + l_{22}w \end{cases}$$

系数  $l_{11}$ 、 $l_{12}$ 、 $l_{21}$ 、 $l_{22}$  是可以计算出来的。

如果  $p_1$  代表了观测值变化最大的方向 (即沿该方向观测值方差最大), 而且  $p_2$  和  $p_1$  正交, 则称  $p_1$  为  $h$ 、 $w$  的第一主成分,  $p_2$  称为  $h$ 、 $w$  的第二主成分。这种分析方法称为主成分法。可以看出:

- ① 新变量  $p_1$ 、 $p_2$  是原始变量  $h$ 、 $w$  的线性函数。
- ②  $p_1$  与  $p_2$  相互垂直, 即两个新变量不相关。

由此推广到一般情况, 实测变量  $x_1 \sim x_m$ , 共测得  $n$  个观测, 数据如表 14-2 所示。

表 14-2 参与因子分析的观测量与变量数据

变量 $j$ 观测量 $i$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	...	$x_m$
1	$x_{11}$	$x_{12}$	$x_{13}$	$x_{14}$	$x_{15}$	...	$x_{1m}$
2	$x_{21}$	$x_{22}$	$x_{23}$	$x_{24}$	$x_{25}$	...	$x_{2m}$
3	$x_{31}$	$x_{32}$	$x_{33}$	$x_{34}$	$x_{35}$	...	$x_{3m}$
4	$x_{41}$	$x_{42}$	$x_{43}$	$x_{44}$	$x_{45}$	...	$x_{4m}$
5	$x_{51}$	$x_{52}$	$x_{53}$	$x_{54}$	$x_{55}$	...	$x_{5m}$
...	...	...	...	...	...	...	...
$n$	$x_{n1}$	$x_{n2}$	$x_{n3}$	$x_{n4}$	$x_{n5}$	...	$x_{nm}$

在原始变量的  $m$  维空间中，找到新的  $m$  个坐标轴，新变量与原始变量的关系可以表示为

$$\begin{cases} p_1 = l_{11}x_1 + l_{12}x_2 + l_{13}x_3 + \Lambda + l_{1m}x_m \\ p_2 = l_{21}x_1 + l_{22}x_2 + l_{23}x_3 + \Lambda + l_{2m}x_m \\ p_3 = l_{31}x_1 + l_{32}x_2 + l_{33}x_3 + \Lambda + l_{3m}x_m \\ \Lambda \ \Lambda \ \Lambda \ \Lambda \ \Lambda \ \Lambda \ \Lambda \ \Lambda \ \Lambda \ \Lambda \ \Lambda \\ p_m = l_{m1}x_1 + l_{m2}x_2 + l_{m3}x_3 + \Lambda + l_{mm}x_m \end{cases}$$

这  $m$  个新变量中可以找到  $l$  个新变量 ( $l < m$ ) 能解释原始数据大部分方差所包含的信息，包含的信息量是原始数据包含信息量的绝大部分。其余  $m-l$  个新变量对方差影响很小。我们称这  $m$  个新变量为原始变量的主成分，每个新变量均为原始变量的线性组合。

(2) 主成分分析中的统计量

前面说到的使得方差最大的  $l$  个互相正交的方向及沿这些方向的方差是一个特征值问题的特征向量和特征值。这些特征值和特征向量为特征方程  $Ax=\lambda x$  的解，这里  $A$  为样本协方差阵或样本相关阵。如果用样本相关阵，可以避免由于各变量量纲不同而产生的问题。如果用样本协方差阵，应该对原始变量进行标准化，这在 SPSS 中是自动完成的。主成分分析的主要统计量如表 14-3 所示。

① 特征方程的根，通常用  $\lambda$  表示。有  $m$  个变量，就有  $m$  个特征方程的根。它是确定主成分数目的根据。SPSS 软件输出列出的特征方程的根已经是经过重新排序重新命名的结果。最大的为  $\lambda_1$ ，最小的为  $\lambda_m$ 。

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_m$$

该统计量反映的是原始变量的总方差在各成分上重新分配的结果。

表 14-3 主成分分析中的主要统计量

成分号 $i$	特征值 $\lambda_i$	贡献率 $\lambda_i/m$	累计贡献率	特征向量 $L_i: l_{i1} l_{i2} \cdots l_{im}$
1	$\lambda_1$	$\lambda_1/m$	$\lambda_1/m$	$L_1: l_{11} l_{12} \cdots l_{1m}$
2	$\lambda_2$	$\lambda_2/m$	$(\lambda_1 + \lambda_2)/m$	$L_2: l_{21} l_{22} \cdots l_{2m}$
3	$\lambda_3$	$\lambda_3/m$	$(\lambda_1 + \lambda_2 + \lambda_3)/m$	$L_3: l_{31} l_{32} \cdots l_{3m}$
...	...	...	...	...
$m$	$\lambda_m$	$\lambda_m/m$	$m$	$L_m: l_{m1} l_{m2} \cdots l_{mm}$

根据方差的定义，第  $i$  个主成分的方差是总方差在各主成分上重新分配后，在第  $i$  个成分上分配的结果，在数值上等于第  $i$  个特征值。

$$S_{p_i} = \frac{\sum_{i=1}^m (p_i - \bar{p}_i)^2}{n-1} = \lambda_i$$

$\sum_{i=1}^m \lambda_i = m$  原始变量个数  $m$  等于特征值的数目  $m$ ， $m$  个特征值之方差总和等于  $m$  个

特征值之和, 等于  $m$ , 即等于标准化的原始变量的方差之总和。

② 各成分之贡献率定义为: 各成分所包含的信息占总信息的百分比。用方差衡量变量所包含的信息量, 则每个成分所提供方差占总方差 ( $m$ ) 的百分比即该成分的贡献率。

即  $P_i$  的贡献率为

$$\frac{\lambda_i}{\sum_{i=1}^m \lambda_i} = \frac{S_{P_i}}{\sum_{i=1}^m S_{P_i}} = \frac{\lambda_i}{m}$$

③ 前  $k$  个成分的累计贡献率为

$$\sum_{i=1}^k \frac{\lambda_i}{\sum_{i=1}^m \lambda_i} = \sum_{i=1}^k \frac{\lambda_i}{m}$$

通常取累计贡献率大于等于 80% 来确定取前  $k$  个成分作为该研究问题的主成分。

④ 确定取几个成分作为主成分的判定方法有两种:

- 取所有特征值大于 1 的成分作为主成分。
- 根据累计贡献率达到的百分比值确定。例如取累计贡献率达到 80%, 其含义是此前  $l$  个成分 (新变量) 所包含的信息占原始变量包含的总信息的 80%, 其余  $m-l$  个新变量对方差影响很小, 我们认为可以接受, 则取前  $l$  个成分作为主成分。

⑤ 特征向量是各成分表达式中的标准化原始变量的系数向量, 就是各成分的特征向量。得出特征向量, 就可以写出每个成分的表达式。注意, 前面公式中得到的使  $S_{p_i} = \lambda_i$  的各个成分  $p_i$  的系数 ( $l_{i1}, l_{i2}, \dots, l_{im}$ ) 是单位特征向量, 并不是 SPSS 输出中的 Component Matrix 中的系数。而 Component Matrix 中的各个分量的系数为上面单位特征向量乘以相应的特征值的平方根的结果。如果令

$$a_{ij} = \sqrt{\lambda_i} l_{ij} \quad i, j = 1, \Lambda, m$$

那么,  $a_{ij}$  为第  $i$  个成分和第  $j$  个变量的相关系数, 也称为载荷 (loading)。SPSS 中的 Component Plot 即 loading plot 选项就是由 Component Matrix 中各个分量系数点出来的。

⑥ 主成分分数

根据主成分表达式和各观测量中各变量值计算出的成分值, 它与上面关于  $p_i$  公式中 (用  $a_{ij}$  代替  $l_{ij}$ , 并且把变量  $x_j$  标准化之后) 得到的  $p_i$  成比例, 称为该观测量的该成分的分数。该成分第几个主成分就称该值为第几个主成分分数。如果在输出中选了该项, 则在原始数据中会增加对每个观测值所计算的主成分的分数。

## 2. 因子分析的概念

### (1) 什么是因子分析

探讨存在相关关系的变量之间, 是否存在不能直接观察到但对可观测变量的变化起支配作用的潜在因子的分析方法称为因子分析。因子分析就是寻找潜在的起支配作用的

因子模型的方法。

设有原始变量:  $x_1, x_2, x_3, \dots, x_m$ 。它们与潜在因子之间的关系可以表示为下式

$$\begin{cases} x_1 = b_{11}z_1 + b_{12}z_2 + b_{13}z_3 + \Lambda + b_{1m}z_m + e_1 \\ x_2 = b_{21}z_1 + b_{22}z_2 + b_{23}z_3 + \Lambda + b_{2m}z_m + e_2 \\ x_3 = b_{31}z_1 + b_{32}z_2 + b_{33}z_3 + \Lambda + b_{3m}z_m + e_3 \\ \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \\ x_m = b_{m1}z_1 + b_{m2}z_2 + b_{m3}z_3 + \Lambda + b_{mm}z_m + e_m \end{cases}$$

其中  $z_1 \sim z_m$  为  $m$  个潜在因子, 是各原始变量都包含的因子, 称共性因子;  $e_1 \sim e_m$  为  $m$  个只包含在某个原始变量之中的, 只对一个原始变量起作用的个性因子, 是各变量特有的特殊因子。

共性因子与特殊因子相互独立。找出共性因子是因子分析的主要目的。计算出结果后要对共性因子的实际含义进行探讨, 并给以命名。

进行因子分析的方法很多, 常用的方法是主成分法。如果特殊因子可以忽略, 可以使用主成分分析的计算方法进行因子分析。

## (2) 因子分析中的统计量

### ① 因子与因子载荷

根据累计贡献率尽量大的原则决定公因子数。公因子数为  $k$ , 初始因子模型为

$$\begin{cases} x'_1 = \alpha_{11}f_1 + \alpha_{12}f_2 + \Lambda + \alpha_{1k}f_k + e_1 \\ x'_2 = \alpha_{21}f_1 + \alpha_{22}f_2 + \Lambda + \alpha_{2k}f_k + e_2 \\ x'_3 = \alpha_{31}f_1 + \alpha_{32}f_2 + \Lambda + \alpha_{3k}f_k + e_3 \\ \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \quad \Lambda \\ x'_m = \alpha_{m1}f_1 + \alpha_{m2}f_2 + \Lambda + \alpha_{mk}f_k + e_m \end{cases}$$

其中  $x'_1 \sim x'_m$  是对原始变量进行均值为 0, 标准差为 1 标准化后的变量。 $f_i$  为第  $i$  个因子,  $\alpha_{ij}$  为  $x'_i$  在共性因子  $f_j$  上的载荷, 它的统计意义就是第  $i$  个变量与第  $j$  个公共因子的相关系数, 表示  $x_i$  依赖  $f_j$  的份量, 载荷的 SPSS 输出是在 Component Matrix 中, 旋转后的载荷在 Rotated Component Matrix 中。

### ② 共性方差

因为  $x'_1 \sim x'_m$  是原始变量  $x_1 \sim x_m$  标准化后的变量, 因此每个变量的方差均为 1。即  $\text{Variance}(x'_i) = 1$ , 记做  $Va(x'_i)$ 。

$$Va(x'_i) = \alpha_{i1}^2 + \alpha_{i2}^2 + \alpha_{i3}^2 + \Lambda + \alpha_{im}^2 + V(e_i) = 1$$

它由两部分组成:

- 一部分是几个共性因子共同引起的共性方差

$$\alpha_{i1}^2 + \alpha_{i2}^2 + \alpha_{i3}^2 + \Lambda + \alpha_{im}^2$$

- 一部分是由特殊因子引起的特性方差  $V(e)$ 。共性方差占总方差的百分比越大, 说

明共性因子的作用越大。因为每个变量的方差均为 1，因此共性方差数值就是所占的百分比数值。

要根据因子载荷和共性方差的大小解释共性因子  $f_i$  的意义，要计算共性方差

$$V \text{ common: } Vc(x'_i) = \sum_{j=1}^m \alpha_{ij}^2$$

如果取前  $k$  个因子，共性方差为

$$Vc(x'_i) = \sum_{j=1}^k \alpha_{ij}^2$$

### ③ 因子得分

因子得分就是每个观测量的共性因子的值。要计算因子得分必须写出共性因子表达式。而共性因子不是能直接观测得到的，它是潜在的，但是可以通过可观测的变量获得。即可以把共性因子表达成可观测变量的线性组合形式，通常用回归方法解决。这样就可以通过每个观测量的各变量值，计算该观测量的因子得分。

### ④ 关于旋转

要结合专业知识解释共性因子具有的实际意义并不是很容易的事。常常得不到满意的解释。数学可以证明，满足模型要求的共性因子并不唯一。只要对初始共性因子进行旋转，就可以获得一组新的共性因子。所谓旋转就是一种坐标变换。在旋转后的新坐标系中，因子载荷将得到重新分配，使公因子负荷系数向更大（向 1）或更小（向 0）方向变化，因此有可能对潜在因子做专业性解释，对公因子的命名和解释变得更加容易。对初始因子进行旋转的方法很多，通常分为两类：

- 一类是能保证旋转后各共性因子仍然正交称正交旋转。如方差最大正交旋转，就是使共性因子上的相对载荷平方的方差之和达到最大，并保证原共性因子之间的正交性和共性方差总和不变。
- 另一类旋转后不能保证各共性因子之间的正交关系。如斜交旋转。

因子分析的一个重要目的在于对原始变量进行分门别类的综合评价。如果因子分析结果保证了因子之间的正交性（不相关），但对因子不易命名，可以通过对因子模型的正交旋转，保证变换后各因子仍正交，这是比较理想的情况。如果经过正交变换后对公因子仍然不易解释，也可以进行斜交旋转，或许可以得到比较容易解释的结果。

### 3. 因子分析过程的功能

SPSS 使用 FACTOR 过程进行因子分析，见图 14-1。主成分分析是作为因子分析的一种（没有旋转的）方法出现的。可以通过对话框指定因子提取的方法，以及控制因子提取进程的参数；可以指定旋转方法；可以对参与因子分析的变量给出描述统计量，指定输出负荷矩阵的格式；还可以产生新变量，其值是因子得分，并将其保存在数据文件中。使用 FACTOR 过程的命令语句和一系列子命令还允许：

- (1) 一个命令完成多种方法的分析，对一种因子提取结果进行多种旋转。
- (2) 指定在提取因子与旋转时进行迭代的收敛判据，控制因子提取及旋转的进程。

- (3) 指定产生单个的旋转因子散点图。
- (4) 具体指定保存多少个因子。
- (5) 把相关矩阵或因子负荷矩阵写到磁盘上，以便进一步分析。
- (6) 指定主轴因子法的对角线上的值。
- (7) 从存储设备读取相关矩阵或因子负荷矩阵，并进一步分析。

4. 因子分析对变量的要求与假设

(1) 在因子分析中研究的是包含原始变量绝大部分信息的综合变量，对原始变量不分因变量和自变量。因子分析要求参与分析的变量必须是等间隔测度的或是比率的数值型变量。分类变量不适合做因子分析。那些明显可以做皮尔逊相关系数计算的数据才适合进行因子分析。观测量应该彼此独立。一般观测量数应该为变量数的 5 倍以上才好。

(2) 因子分析的前提。因子分析模型指定变量由公因子（由模型估计的因子）和特殊因子（与原始观测变量不交迭）确定。参数计算的前提是，假设所有特殊因子彼此不相关，而且与公因子也不相关。

14.1.2 因子分析过程

对于初学统计分析的读者，可以完全使用系统默认值进行最简单的因子分析。虽然可能得不到非常满意的结果，但通过初步分析可以对所研究的问题有一个初步的认识，对进一步的分析会有帮助。对比较简单的问题，有时只使用系统默认值进行因子分析就可以得到比较满意的结果。

【例 1】data14-01 中的数据是美国洛杉矶标准大城市统计区中的 12 个人口调查区的 5 个经济学指标（变量）的数据。以对 12 个地区的 5 个经济指标的调查数据进行因子分析为例，说明因子分析过程。

1. 定义变量及标签：no（编号）、pop（总人口）、school（中等学校平均校龄）、employ（总雇员数）、services（专业服务项目）、house（中等房价）。

2. 使用默认值进行因子分析

(1) 读取数据文件 data14-01。按 Analyze→Data Reduction→Factor 顺序单击菜单项，

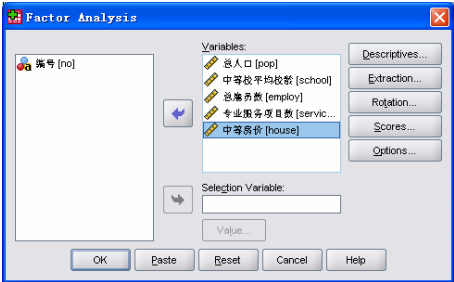


图 14-3 因子分析主对话框

展开 Factor Analysis 对话框，如图 14-3 所示。

- (2) 指定参与分析的变量。
- (3) 在源变量框中选择变量，把选中的变量名移到右面的 Variables 框中。本例将 pop、school、employ、services、house 五个变量移到 Variables 框中。
- (4) 单击 OK 按钮，运行 FACTOR 过程。
- (5) 输出结果见表 14-4～表 14-6。

表 14-4 为公因子提取前与公因子提取后的公因子方差表。

**Initial** 是在提取因子（或成分，系统默认的是主成分法）之前的各变量的公因子方差。对主成分分析来说，该值是要被分析的矩阵（相关矩阵或协方差矩阵）的对角元素。对因子分析来说，这些值是用其他变量作为预测变量时每个变量的载荷的平方和。由于分析的是相关阵，原始变量的公因子方差均为 1（如果分析的是协方差阵，此处为各变量的方差），五个变量的公因子方差之总和为 5。

表 14-4 公因子方差表

Communalities		
	Initial	Extraction
总人口	1.000	.988
中等校平均校龄	1.000	.885
总雇员数	1.000	.979
专业服务项目数	1.000	.880
中等房价	1.000	.938

Extraction Method: Principal Component Analysis.

**Extraction** 是各变量的未旋转的公因子方差。这些公因子方差是用作预测因子变量的多重相关的平方。表中的公因子方差都很高，它表明提取的成分能很好地描述这些变量。

表 14-5 总方差分解

表 14-6 主成分分析的因子载荷阵

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.873	57.466	57.466	2.873	57.466	57.466
2	1.797	35.933	93.399	1.797	35.933	93.399
3	.215	4.297	97.696			
4	.100	1.999	99.695			
5	.015	.305	100.000			

Extraction Method: Principal Component Analysis.

	Component	
	1	2
总人口	.581	.806
中等校平均校龄	.767	-.545
总雇员数	.672	.726
专业服务项目数	.932	-.104
中等房价	.791	-.558

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

表 14-5 为各成分的公因子方差表。

其中：**Components** 为各成分的序号。

**Initial Eigenvalues** 是相关矩阵或协方差矩阵的特征值。这些值是用于确定哪些因子（或成分）应保留，共有三项：

- **Total** 各成分的特征值。第一成分特征值为 **Total=2.873**，第二成分特征值为 **Total=1.797**。本例只有前两个因子的特征值大于 1。
- **% of Variance** 各成分所解释的方差占总方差的百分比。也就是各因子特征值占特征值总和的百分比。
- **Cumulative %** 自上至下各因子方差占总方差百分比的累积百分比。前两个因子的特征值之和占总方差的 93.4%。即前两个因子解释原始 5 个变量的 93.4% 的变异。

**Extraction Sums of Squared Loadings** 为因子提取结果，是未经旋转的因子载荷的平方和。它给出的是每个因子（或成分）的特征值说明的方差占总方差的百分比和累计百分比。从初始分析的统计量可以看出按照系统默认值给出的分析原则，提取原则是特征值大于 1，那么应该取前两个因子（就本次分析来说也可称作主成分）。而前两个因子已经对大多数数据给出了充分的概括，可以看出前两个成分所解释的方差占总方差的 93.4%。因此，最后结果是确定提取两个主成分。所以使用这些成分相当大程度上减少了原始数据的复杂性，仅丢失 6.6% 的信息。



表 14-6 为（仅对不做旋转的主成分分析而言）因子载荷阵。它显示了原始变量与各主成分之间的相关程度。根据他们的相关程度的大小，综合出各因子的含义。

可以看出，第一主成分与三个变量的相关较高，这三个变量是专业服务项目、中等校平均校龄和中等房价。而第二成分与总人口数和总雇员数的相关更高些。

由以上输出结果可以认为对因子的提取结果是比较理想的。但是要想对两个因子命名就感到比较困难，每个因子与原始变量相关系数没有很明显的差别。因此为了对因子进行命名，可以进行旋转，使系数向 0 和 1 两极分化。这就要使用选项了。

3. 因子分析过程选项

(1) 在主对话框中，单击 Descriptives 按钮，展开如图 14-4 所示的对话框，从中选择描述统计量。

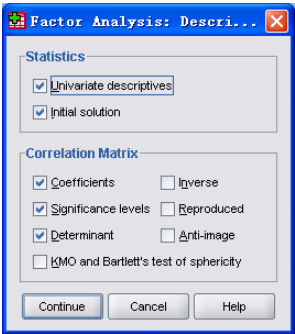


图 14-4 描述统计量对话框

- ① Statistics 统计量栏
  - Univariate descriptives，输出参与分析的原始变量的均值、标准差等单变量描述统计量。
  - Initial solution，给出因子提取前，分析变量的公因子方差。对主成分分析来说，这些值是分析变量的相关或协方差矩阵的对角元素。对因子分析来说，是每个变量用其他变量作预测因子的载荷平方和。
- ② Correlation Matrix 相关矩阵栏。
  - Coefficients，原始分析变量间的相关系数矩阵。
  - Significance levels，每个相关系数相对于 0 的单尾假设

检验的显著性水平。

- Determinant，相关系数矩阵的行列式。
- Inverse，相关系数矩阵的逆矩阵。
- Reproduced，再生相关矩阵。此项给出因子分析后的相关矩阵，还给出残差，即原始相关与再生相关之间的差值。
- Anti-image，给出反映像相关矩阵，包括偏相关系数的负数；反映像协方差矩阵，包括偏协方差的负数。在一个好的因子模型中除对角线上的系数较大外，远离对角线上的元素的系数应该比较小。
- KMO and Bartlett's test of sphericity，要求进行 KMO 检验和球形 Bartlett 检验。选择此项给出对采样充足度的 Kaisex-Meyer-Olkin 测度，检验变量间的偏相关是否很小。Bartlett 球形检验，检验相关矩阵是否是单位矩阵，它表明因子模型是否是不合适宜的。也就是说，你的数据是否适合做因子分析。

(2) 在主对话框中，单击 Extraction 按钮，展开如图 14-5 所示的对话框。

① 因子提取方法选项

Method 参数框是下拉列表给出一组提取方法的选项。提供七种提取方法供选择：

- **Principal components**, 主成分法。该方法假设变量是因子的纯线性组合。第一成分有最大的方差, 后续的成分, 其可解释的方差逐个递减。主成分法是常用的获取初始因子分析结果的方法。它假设特殊因子作用可以忽略不计。

- **Unweighted least square**, 不加权最小平方方法。该方法使观测的和再生相关矩阵之差的平方和最小, 不计对角元素。

- **Generalized least squares**, 用变量值的倒数加权, 使观测的和再生的相关矩阵之差的平方和最小。给较高值的权重比给较低值的权重要小。

- **Maximum likelihood**, 最大似然法。此方法不要求多元正态分布。该方法给出参数估计。如果样本来自多元正态总体, 它们与原始变量的相关矩阵极为相似。用变量单值倒数对原始分析变量加权。

- **Principal axis factoring**, 使用多元相关的平方作为对公因子方差的初始估计。初始估计公因子方差时, 多元相关系数的平方置于对角线上。这些因子载荷用于估计新公因子方差, 替换对角线上的前一次的公因子方差估计。每次迭代结束都计算从上次到本次迭代结果公因子方差的变化量。这样的迭代持续到公因子方差的变化量满足提取因子的收敛判据时为止。

- **Alpha factoring**,  $\alpha$  因子提取法。
- **Image factoring**, 映像因子提取法 (由 Guttman 提出)。根据映像学原理提取公因子, 并把一个变量看作其他各变量的多元回归, 而不是假设因子的函数。

② 在 **Analyze** 栏中指定分析矩阵的选项。

- **Correlation matrix**, 使用变量的相关矩阵进行提取因子的分析。如果参与分析的变量的测度单位不同时, 应该选择此项。

- **Covariance matrix**, 使用变量的协方差矩阵进行提取因子的分析。如果参与分析的变量测度单位相同, 可以选择此项。

③ 在 **Extract** 栏中, 选择提取结果。理论上因子数目与原始变量数目相等, 但因子分析的目的是用少量因子代替多个原始变量。选择提取多少个因子由本组选项决定。

- **Eigenvalues over**, 指定提取的因子应该具有的特征值范围, 在此项后面的矩形框中给出, 系统默认值为 1, 即要求提取那些特征值大于 1 的因子。指定特征值来决定提取因子数目的方法是系统默认的方法。

- **Number of factors**, 指定提取公因子的数目。选择此项后, 将指定的数目输入到该项后面的矩形框中, 数值应该是 0 至分析变量数目之间的正整数。

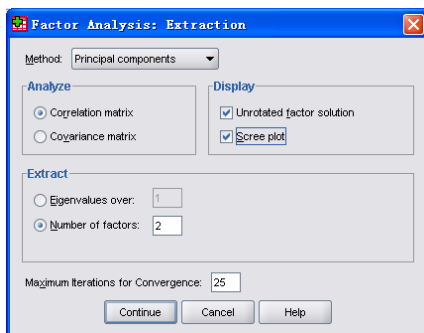


图 14-5 因子提取子对话框

④ 在 Display 栏中指定与因子提取有关的输出项。

- Unrotated factor solution, 要求显示未经旋转的因子提取结果。此为系统默认的。
- Scree plot, 要求显示按特征值大小排列因子, 以特征值为两个坐标轴绘制碎石图。

该图有助于确定保留多少个因子。典型的碎石图会有一个明显的拐点, 在该点之前是与大因子有关的陡峭的折线, 之后是与小因子有关的缓坡折线。

⑤ Maximum iterations for Convergence 参数框, 指定因子分析停止的最大迭代次数, 系统默认的最大迭代次数为 25。可以修改该值。

(3) 在主对话框中, 单击 Rotation 按钮, 在如图 14-6 所示的对话框中选择旋转方法。

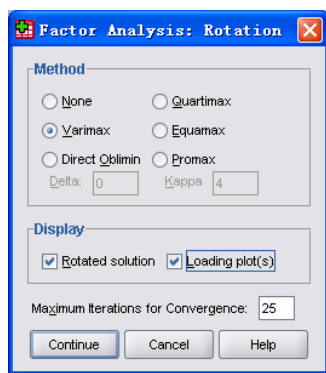


图 14-6 旋转方法选择对话框

① 在 Method 栏中选择旋转方法

- None, 不进行旋转。是系统默认的选项。
- Varimax, 方差最大旋转, 是一种正交旋转。它使每个因子上的具有最高载荷的变量数最小, 因此可以简化对因子的解释。

• Direct Oblimin, 直接斜交旋转, 指定此项可以在下面的矩形框中输入  $\delta$  值, 0 值产生最高相关因子。 $\delta$  值越接近 0, 斜交程度越深。大负数产生旋转的结果与正交接近。要不想  $\delta$  值为 0, 可以输入一个小于等于 0.8 的值。

• Quartimax, 四次最大正交旋转。该旋转方法使每个变量中需要解释的因子数最少。可以简化对变量的解释。

• Equamax, 平均正交旋转, 是简化对因子解释的 Varimax 方法与简化对变量解释的 Quartimax 方法的结合。可以使在一个因子上有高载荷的变量数和变量中需要解释的因子数最少。

• Promax, 斜交旋转方法, 允许因子彼此相关。它比直接斜交旋转更快, 因此适用于大数据集的因子分析。

② 在 Display 栏中选择有关输出的选项

• Rotated solution, 旋转结果。指定旋转方法才能指定此项。指定此项将对正交旋转显示旋转后的因子矩阵、因子转换矩阵, 对斜交旋转显示旋转后的因子矩阵、因子结构矩阵和因子间的相关矩阵。

• Loading plot(s), 因子载荷散点图。指定此项将给出以两两因子为坐标轴的各变量的载荷散点图。如果有两个因子, 给出各原始变量基于 Rotated Component Matrix 表输出数据的散点图; 如果多于两个因子, 则给出前三个因子的三维因子载荷散点图; 如果只提取了一个因子, 则不会输出载荷散点图。注意, 选择此项给出的是经旋转后的因子载荷图。

③ Maximum Iterations for Convergence 参数框, 指定旋转收敛的最大迭代次数, 系统默认值为 25, 可以在此项后面的矩形框中输入指定值。

(4) 在对话框中, 单击 **Scores** 按钮, 展开 **Factor Scores** 对话框, 如图 14-7 所示。

① **Save as variables**, 将因子得分作为新变量保存在数据文件中。每次分析产生一组新变量, 每次分析产生多少个因子, 就生成多少个新变量。新变量名的最后一个数字表示分析的顺序号。因子序号占倒数第三个字符的位置, 倒数第二个字符为“-”。在输出窗口中给出对因子得分的命名和变量标签, 表明用以计算因子得分的方法。

② 在 **Method** 栏中指定计算因子得分的方法

选择 **Save as variables** 项, 激活 **Method** 栏中各项。

- **Regression**, 回归法。其因子得分的均值为 0, 方差等于估计因子得分与实际因子得分之间的多元相关的平方。

- **Bartlett**, 巴特利特法。因子得分均值为 0。超出变量范围的特殊因子平方和被最小化。

- **Anderson-Rubin**, 安德森—鲁宾法。是为了保证因子的正交性而对巴特利特因子得分的调整。其因子得分的均值为 0, 标准差为 1, 且彼此不相关。

③ **Display factor score coefficient matrix**, 在输出窗口中显示因子得分系数矩阵, 是标准化的得分系数。原始变量值进行标准化后, 可以根据该矩阵给出的系数计算各观测量的因子得分。还显示协方差矩阵。

(5) 在主对话框中, 单击 **Options** 按钮, 展开如图 14-8 所示的对话框。

① 在 **Missing Values** 栏中选择处理缺失值方法

- **Exclude cases listwise**, 在分析过程中对指定的分析变量中有缺失值的观测量一律剔除。所有分析变量带有缺失值的观测量都不参与分析。

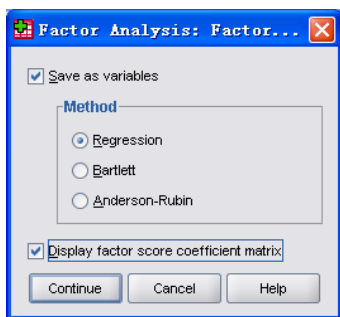


图 14-7 因子得分选项对话框

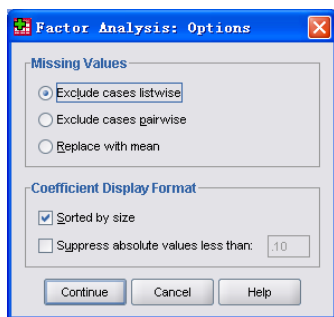


图 14-8 选择输出项对话框

- **Exclude cases pairwise**, 成对剔除带有缺失值的观测量。即在计算两个变量的相关系数时, 只把这两个变量中带有缺失值的观测量剔除。选择此项可以最大限度利用得来不易的原始数据。

- **Replace with mean**, 用变量的均值代替该变量的所有缺失值。

② 在 **Coefficient Display Format** 栏中决定载荷系数的显示格式

- **Sorted by size**, 载荷系数按其数值的大小排列并构成矩阵, 使在同一因子上具有较高载荷的变量排在一起, 便于得出结论。

- **Suppress absolute values less than**, 不显示那些绝对值小于指定值的载荷系数。要在该项的框中输入 0~1 之间的数作为临界值, 系统默认的临界值为 0.10。选择此项可以突出载荷较大的变量, 便于得出结论。

### 14.1.3 因子分析实例

【例 2】仍使用美国洛杉矶对 12 个人口调查区调查的数据进行因子分析。

#### 1. 操作步骤

(1) 读取数据文件 data14-01。按 **Analyze**→**Data Reduction**→**Factor** 顺序单击菜单项, 展开 **Factor Analysis** 对话框。

(2) 将 **pop**、**school**、**employ**、**services**、**house** 五个变量移到 **Variables** 框中。

(3) 在主对话框中, 单击 **Descriptives** 按钮, 展开相应的对话框。

① 在 **Statistics** 栏中选择要求输出的统计量

- 选中 **Univariate Descriptives**, 要求显示单变量的描述统计量。

- 选中 **Initial solution**, 要求显示初始因子分析结果。

② 在 **Correlation Matrix** 栏中选择要求输出的相关矩阵

- 选中 **Coefficients**, 要求显示相关矩阵的相关系数。

- 选中 **Significance levels**, 要求显示针对相关系数为 0 的假设检验显著性概率。

(4) 在主对话框中, 单击 **Extraction** 按钮, 展开 **Extraction** 对话框。

① 在 **Method** 因子提取方法参数框, 选择 **Principal components** 主成分分析选项。

② 在 **Analyze** 栏中选择 **Correlation matrix** 分析相关矩阵项。

③ 在 **Extract** 栏中选择 **Number of factors**, 并在其小矩形框中输入提取因子数 2。

④ 在 **Display** 栏中选择要求的输出项:

- 选中 **Unrotated factor solution**, 在输出窗口中显示旋转前的因子分析结果。

- 选中 **Scree plot**, 在图表窗口中显示因子碎石图。

⑤ 在 **Maximum Iterations for Convergence** 参数框中, 选择停止迭代的最大迭代次数。使用默认值 25。

(5) 在主对话框中单击 **Rotation** 按钮, 展开 **Rotation** 对话框, 见图 14-6。

① 在 **Method** 旋转方法栏中, 选择 **Varimax** 最大方差旋转。

② 在 **Display** 栏中选择 **Rotated solution** 和 **Loading plot(s)**, 前者要求显示旋转后的结果, 后者要求显示因子载荷图。

(6) 在主对话框中, 单击 **Scores** 按钮, 展开 **Factor Scores** 对话框, 见图 14-7。

① 选中 **Save as variables**, 以变量形式将因子得分保存在数据文件中, 使用 **Method** 栏中默认的 **Regression**。

- ② 选中 Display factor score coefficient matrix, 输出因子得分系数矩阵。
- (7) 在主对话框中, 单击 Options 按钮, 展开 Options 对话框, 见图 14-8。
- ① 在 Missing Value 栏中选择 Exclude cases Listwise。
- ② 在 Coefficient Display Format 栏中选择 Sorted by size。
- (8) 在主对话框中, 单击 OK 按钮执行运算。单击 Paste 按钮, 在 Syntax 窗口中生成相应的命令语句。

## 2. 执行如下程序

```

FACTOR                                ①
/VARIABLES employ house pop school services /MISSING LISTWISE /ANALYSIS
      employ house pop school services                                ②
/PRINT UNIVARIATE INITIAL CORRELATION SIG EXTRACTION
      ROTATION FSCORE                                              ③
/FORMAT SORT                                                         ④
/PLOT EIGEN ROTATION                                                ⑤
/CRITERIA FACTORS(2) ITERATE(25) /EXTRACTION PC                    ⑥
/CRITERIA ITERATE(25)                                              ⑦
/ROTATION VARIMAX                                                  ⑧
/METHOD=CORRELATION                                             ⑨
/SAVE REG(ALL).                                                    ⑩

```

## 3. 程序语句解释如下:

- ① FACTOR 语句调用 FACTOR 过程。
- ② 三个语句 VARIABLES、MISSING 和 ANALYSIS 分别列出变量表、指定分析变量和缺失值处理方法。
- ③ PRINT 语句指定在输出窗口显示的内容有单变量描述统计量、初始分析结果、相关矩阵及显著性检验结果、因子提取结果、旋转后的结果和因子得分。
- ④ 指定显示因子表达式时各原始变量按负荷大小顺序显示。
- ⑤ 要求按因子特征值绘制碎石图和旋转后的因子载荷图。
- ⑥ 指定提取 2 个公因子, 迭代次数要求小于等于 25。EXTRACTION 语句要求使用主成分法提取因子。
- ⑦ 旋转的迭代次数选择小于等于 25。
- ⑧ 旋转方法要求使用最大方差法。
- ⑨ 使用相关矩阵进行分析计算。
- ⑩ 使用回归方法计算因子得分, 并全部作为新变量保存在数据文件中。

## 4. 执行结果见表 14-7~表 14-13, 图 14-9~图 14-11。

此处略去公因子方差表和各因子方差分解表, 分别与表 14-4 和表 14-5 相同。

5. 结果解释、分析与结论。

表 14-7 为单变量描述统计量。自左至右显示了变量标签、各变量的均值、各变量的标准差、参与计算这些统计量的观测量数。

表 14-8 为各分析变量的相关矩阵。

表 14-7 单变量描述统计量

Descriptive Statistics			
	Mean	Std. Deviation	Analysis N
总人口	6241.67	3439.994	12
中等校平均校龄	11.442	1.7865	12
总雇员数	2333.33	1241.212	12
专业服务项目数	120.83	114.928	12
中等房价	17000.00	6367.531	12

表 14-8 原始变量的相关矩阵

Correlation Matrix						
	总人口	中等校平均校龄	总雇员数	专业服务项目数	中等房价	
Correlation	总人口	.010	.972	.439	.022	
	中等校平均校龄	1.000	.154	.691	.863	
	总雇员数	.972	1.000	.515	.122	
	专业服务项目数	.439	.691	1.000	.778	
	中等房价	.022	.863	.122	1.000	
Sig. (1-tailed)	总人口	.488	.000	.077	.472	
	中等校平均校龄	.488	.316	.006	.000	
	总雇员数	.000	.316	.043	.353	
	专业服务项目数	.077	.006	.043	.001	
	中等房价	.472	.000	.353	.001	

图 14-9 表现各成分特征值的碎石图。分析碎石图可以看出因子 1 与因子 2，以及因子 2 与因子 3 之间的特征值之差值比较大。而因子 3、4、5 之间的特征值差值都比较小。可以初步得出保留两个因子将能概括绝大部分信息。明显的拐点为 3，因此提取 2 个因子比较合适。证实了表 14-5 中的结果。

表 14-9 是初始提取的因子载荷矩阵。相关系数比较接近，不好命名。

表 14-9 旋转前因子载荷矩阵

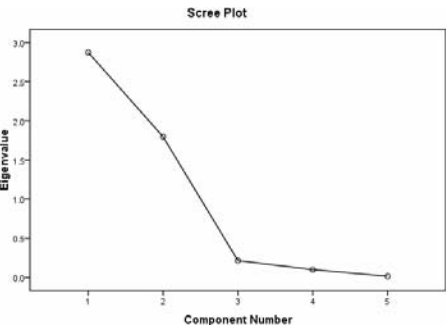


图 14-9 特征值碎石图

Component Matrix <sup>a</sup>		
	Component	
	1	2
专业服务项目数	.932	-.104
中等房价	.791	-.558
中等校平均校龄	.767	-.545
总人口	.581	.806
总雇员数	.672	.726

Extraction Method: Principal Component Analysis.  
a. 2 components extracted.

表 14-10 为因子旋转的转换矩阵。

表 14-11 是旋转后因子载荷矩阵。表下方是有关因子提取与旋转方法的说明：使用主成分法提取因子，使用 Varimax 最大方差法旋转，经 3 次迭代收敛。

表 14-10 转换矩阵

Component Transformation Matrix		
Component	1	2
1	.821	.571
2	-.571	.821

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.

表 14-11 旋转后因子载荷矩阵

Rotated Component Matrix <sup>a</sup>		
	Component	
	1	2
中等房价	.968	-.006
中等校平均校龄	.941	-.009
专业服务项目数	.825	.447
总人口	.016	.994
总雇员数	.137	.980

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.  
a. Rotation converged in 3 iterations.

表中给出了旋转后的因子（或成分）与原始变量的相关矩阵，是按系数由大到小排列的。可以看出经过旋转后相关系数已经明显地发生了变化了。第一个主成分 Component 1 对中等房价 house、中等校平均校龄 school、专业服务项目 services 有绝对值较大的相关系数，第二个因子相关系数绝对值较大的正好是五个原始变量中的另外两个，即总人口 pop 和总雇员数 employ。根据这些变量的原始含义可以对两个因子进行命名。第一个因子主要概括了一般的社会福利情况的因子，中等房价、中等校校龄和社会服务项目数可以命名为福利条件因子。第二个因子主要概括了人的情况，人口数和就业人数，可以称为人口因子。

表 14-12 为因子得分系数矩阵。根据因子得分系数和原始变量的标准化值，可以计算每个观测量的各因子的得分数，并可以据此对观测量进行进一步的分析。旋转后的因子（主成分）表达式可以写成

$$\text{FAC1}_1 = -0.091 \times \text{pop}' + 0.392 \times \text{school}' - 0.039 \times \text{employ}' + 0.299 \times \text{services}' + 0.403 \times \text{house}'$$

$$\text{FAC2}_1 = 0.484 \times \text{pop}' - 0.096 \times \text{school}' + 0.465 \times \text{employ}' + 0.138 \times \text{services}' - 0.098 \times \text{house}'$$

注意：在因子表达式中的各变量均为经过均值为 0，标准差为 1 标准化后的变量。用原变量名加“'”表示。

表 14-13 是估计回归因子分数的协方差矩阵，即因子（两个主成分）间的相关矩阵。可以看出旋转后 Component 1 与 Component 2 是完全不相关的。这也是因为正交旋转（Varimax）后因子仍然正交。

表 14-12 因子得分系数矩阵

Component Score Coefficient Matrix		
	Component	
	1	2
总人口	-.091	.484
中等校平均校龄	.392	-.096
总雇员数	-.039	.465
专业服务项目数	.299	.138
中等房价	.403	-.098

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.  
Component Scores.

表 14-13 估计回归因子分数的协方差矩阵

Component Score Covariance Matrix		
Component	1	2
1	1.000	.000
2	.000	1.000

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.  
Component Scores.

图 14-10 为旋转后的因子（成分）载荷图，分别以第一主成分和第二主成分为横、纵轴坐标，按表 14-12 中数据作图得到主成分图（图中的指示线是作者加的）。从图中可以看出旋转后各成分的变量更集中了。

图 14-11 是在数据编辑窗口中，以新变量的形式保存的因子得分信息。数据文件中因子分数变量的命名：FAC1\_1 是第一次分析的第一个回归因子分数，FAC2\_1 标签是第 1 次分析的第二个回归因子分数变量。可以将此带有新变量的数据窗口中的数据保存为另一个数据文件 data14-01a。

有了观测量的因子得分变量的值，我们可以进一步对观测量估计因子得分变量进行聚类分析，进一步对每个调查区进行人口与福利方面的分类或分析。



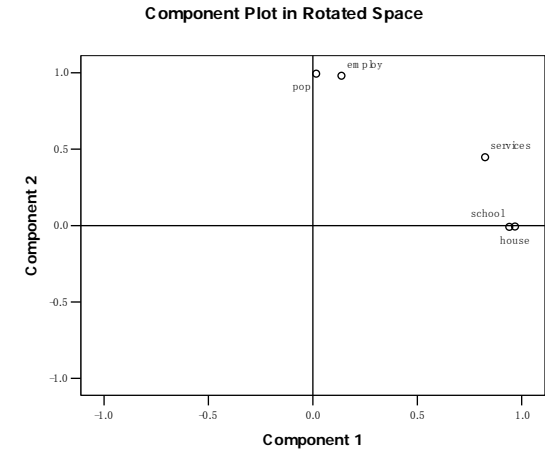


图 14-10 旋转后的因子图

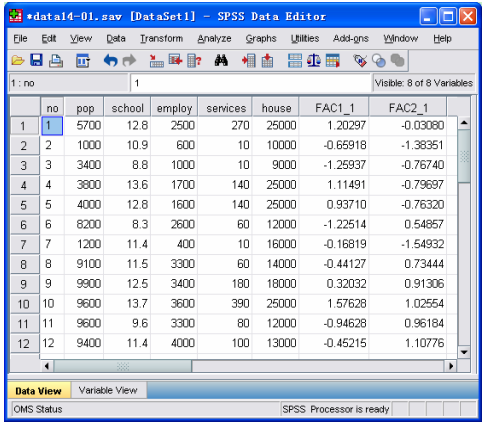


图 14-11 各观测量的两个因子得分的新变量

### 14.1.4 利用因子得分进行聚类

【例 3】下面是利用新变量对 12 个调查区进行聚类的分析的过程及结果：聚类要求聚为两类、三类、四类，然后利用 Graph 功能作散点图，比较分为两类和三类的结果。

#### 1. 操作步骤

(1) 完成因子分析后，在数据文件中保存有各观测量的因子得分，见图 14-11。或直接读取数据文件 data14-01a。该文件保存有各观测量因子得分。

(2) 按 Analyze→Classify→Hierarchical Cluster 顺序单击菜单项，展开 Hierarchical Cluster Analysis 主对话框。

(3) 在主对话框中：

- ① 指定 FAC1\_1 和 FAC2\_1 变量为分析变量，进入 Variable(s)框中。
- ② 指定编号变量 no 作为标识变量，进入 Label Cases By 下面的框中。
- ③ Cluster 栏中选择 Cases，要求进行观测量聚类。
- ④ Display 栏中选择两项 Statistics 和 Plots。

(4) 在主对话框中，单击 Statistics 按钮，展开 Statistics 对话框。选中 Proximity matrix。在 Cluster Membership 栏中选择 Range of Solution，并在 Minimum number of clusters 后面的框中输入 2；在 Maximum number of clusters 后面的框中输入 4，即要求聚为 2 类到 4 类的结果。

(5) 在主对话框中，单击 Plots 按钮，展开 Plots 对话框。在对话框中作如下选择：

- ① 选中 Dendrogram，要求作树形图。

② 在 **Icicle** 栏中指定要在冰柱图中出现的类的范围。选中 **Specified range of clusters** 指定类范围, 在 **Start cluster** 后面的框中输入起始类数 2、在 **Stop cluster** 后面的框中输入终止类数 4; 在 **By** 后面的框中输入步长 1。

③ 在 **Orientation** 栏中选择 **Vertical**。要求显示方向是纵向的。

(6) 在主对话框中, 单击 **Method** 按钮, 展开 **Method** 对话框。

① 在 **Cluster Method** 参数框中选择 **Between-groups linkage**。

② 在 **Measure** 栏中的 **Interval** 后的下拉列表中选择 **Squared Euclidean Distance**, 要求根据两个因子间的欧氏距离的平方进行聚类。

③ 在 **Transform Values** 栏中的 **Standardize** 下拉列表中选择默认值 **None**, 因为两个因子得分本身就是根据标准化变量得出的无量纲变量。

(7) 在主对话框中单击 **Save** 按钮, 展开 **Save New Variables** 对话框。在 **Cluster membership** 栏中选择 **Range of solution**, 并在 **Minimum number of clusters** 后面的框中输入 2; 在 **Maximum number of clusters** 后面的框中输入 4, 即要求保存 3 个新变量, 表示聚为 2、3、4 类时每个观测变量各归为哪一类。

通过作散点图观察 12 个调查区的经济情况分析。对各选项的含义请参考第 11 章的有关内容。

(8) 在主对话框中, 单击 **OK** 按钮, 执行运算或单击 **Paste** 按钮, 得到运行的程序。

2. 运行的程序为:

```
CLUSTER FAC1_1 FAC2_1
```

```
/METHOD BAVERAGE /MEASURE= SEUCLID/ID=no /PRINT SCHEDULE CLUSTER(2,4)
```

```
/PRINT DISTANCE /PLOT DENDROGRAM VICICLE(2,4,1)
```

```
/SAVE CLUSTER(2,4).
```

对于有关语句的说明请参考第 14 章的有关内容。

3. 输出结果见表 14-14、表 14-15、图 14-12、图 14-13。非关键的输出表没有列出。

表 14-14 相似性矩阵

Proximity Matrix												
Case	Squared Euclidean Distance											
	1: 1	2: 2	3: 3	4: 4	5: 5	6: 6	7: 7	8: 8	9: 9	10: 10	11: 11	12: 12
1: 1	.000	5.297	6.606	.595	.607	6.231	4.186	3.289	1.670	1.255	5.605	4.036
2: 2	5.297	.000	.740	3.491	2.933	4.053	.269	4.533	6.234	10.801	5.583	6.249
3: 3	6.606	.740	.000	5.638	4.825	1.733	1.802	2.925	5.319	11.256	3.088	4.168
4: 4	.595	3.491	5.638	.000	.033	7.286	2.212	4.767	3.556	3.534	7.342	6.084
5: 5	.607	2.933	4.825	.033	.000	6.396	1.840	4.143	3.190	3.608	6.523	5.431
6: 6	6.231	4.053	1.733	7.286	6.396	.000	5.518	.649	2.521	8.075	.249	.910
7: 7	4.186	.269	1.802	2.212	1.840	5.518	.000	5.290	6.302	9.673	6.911	7.141
8: 8	3.289	4.533	2.925	4.767	4.143	.649	5.290	.000	.612	4.155	.307	.139
9: 9	1.670	6.234	5.319	3.556	3.190	2.521	6.302	.612	.000	1.590	1.607	.635
10: 10	1.255	10.801	11.3	3.534	3.608	8.075	9.673	4.155	1.590	.000	6.367	4.121
11: 11	5.605	5.583	3.088	7.342	6.523	.249	6.911	.307	1.607	6.367	.000	.265
12: 12	4.036	6.249	4.168	6.084	5.431	.910	7.141	.139	.635	4.121	.265	.000

This is a dissimilarity matrix

表 14-15 聚为 2、3、4 类的结果

Cluster Membership			
Case	4 Clusters	3 Clusters	2 Clusters
1: 1	1	1	1
2: 2	2	2	2
3: 3	2	2	2
4: 4	1	1	1
5: 5	1	1	1
6: 6	3	3	1
7: 7	2	2	2
8: 8	3	3	1
9: 9	3	3	1
10: 10	4	1	1
11: 11	3	3	1
12: 12	3	3	1

从输出信息很难看出各调查区在经济特性方面的区别。五个变量转变为两个综合指标（两个因子）的好处在于减少了指标数目（降维），而综合指标包含的信息没有损失多少。使用两个综合指标可以对调查区的经济状况更清楚地进行分析，还可以使用其他 SPSS 过程进行进一步分析。

Vertical Icicle																
Number of clusters	Case															
	5:3	7:7	2:2	6:6	12:12	8:8	11:11	6:6	10:10	5:5	4:4	1:1				
2	X	X	X	X												
3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

图 14-12 平均连接法形成的冰柱图

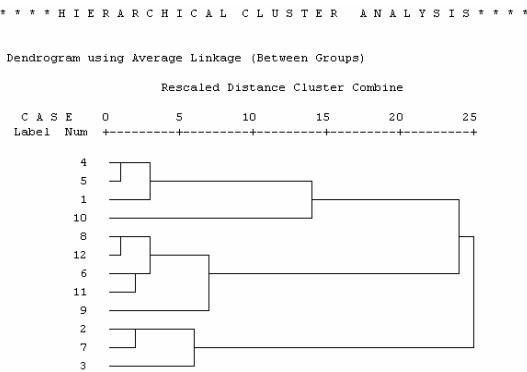


图 14-13 反映聚类全过程的树形图

4. 利用因子得分变量作散点图。
- (1) 按 Graphs→Legacy Dialogs→Scatter/Dot...顺序单击菜单项，打开 Scatter/Dot...对话框，选择 Simple Scatter 项，单击 Define 按钮后，展开 Simple Scatterplot 对话框，如图 14-14 所示。
- (2) 选择因子 2 的得分 FAC2\_1 作为 Y 轴变量送入 Y Axis 中；选择因子 1 的得分 FAC1\_1 作为 X 轴变量送入 X Axis 中；每个观测量用它们的编号标识，故将变量 no 送入 Label cases by 框中；每个观测量按所属类别，使用不同的颜色或符号区分，先作聚为两类的散点图，将变量 Clu2\_1 送入 Set Markets by 框中。按 Clu2\_1（标签为 Average Linkage Between groups）的类数确定符号的类数。
- (3) 单击 Options 按钮，打开相应对话框如图 14-15 所示。选择 Display chart with case labels。注意，选择此项，Label cases by 指定的变量值会标在散点图中观测量点旁。
- (4) 主对话框中单击 OK 按钮，在输出窗中生成的散点图如图 14-16 所示。
- (5) 再把变量 Clu3\_1 移入 Set Markets by 栏，代替 Clu2\_1 得到图 14-17。

注意，图 14-16 和图 14-17 都是经过编辑的图形，为了弥补黑色印刷符号以深浅代替不同颜色，不易观察的问题，改变了其中的分类标识符。读者自己在彩色显示器上可以分辨不同颜色的统一符号标识的分类，不用再编辑。

从图 14-16 可以看出，如果将调查区分为两类，第 2、3、7 区类号为 2 的，是福利因素和人口因素均比较低的；其余调查区的这两个因素水平比较高，可以认为经济状况是相对来说比较好的。

从图 14-17 可以更细致地划分和分析各调查区的经济水平。

① 类号为 2 的调查区有编号为 2、3、7 三个地区，在图的左下角，是两个因子得分均比较低的，可以认为从 5 个经济指标来看均较差的地区。

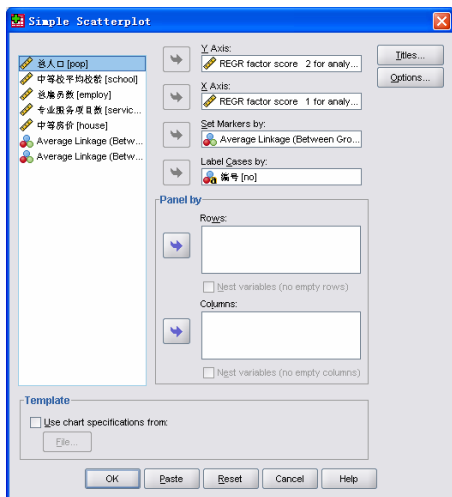


图 14-14 定义散点图坐标系对话框

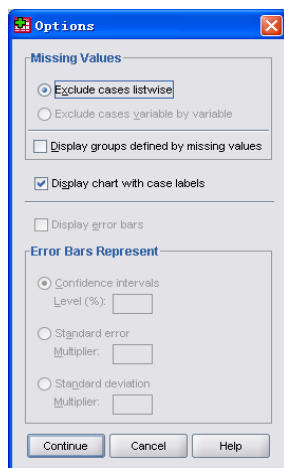


图 14-15 选择项对话框

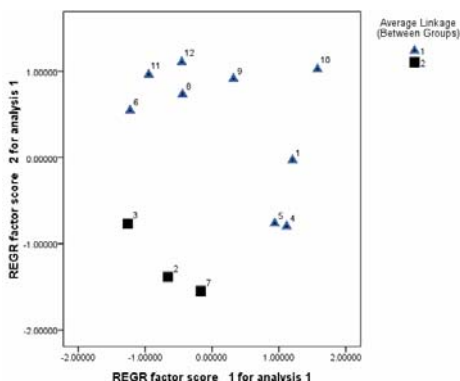


图 14-16 聚为两类的因子得分散点图

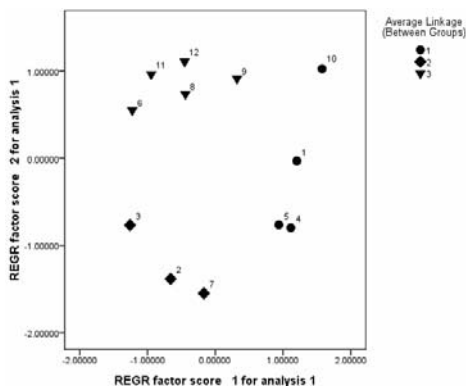


图 14-17 聚为三类的因子得分散点图

② 类号为 3 的调查区 FAC1\_1 比较低，即福利因子得分较低；而 FAC2\_1 比较高，即人口因子得分较高，说明总人口多，就业人数多，但反映福利的学校、服务项目、中等房价均比较低。这样的地区有 6、8、9、11、12 号地区。

③ 类号为 1 的调查区位于散点图的右偏上方，可以看作人口和就业人数均较少、福利条件比较好的地区，有编号为 1、4、5 号地区。

④ 如果分为 4 类，则右上角的点将单独分为一类，是两个因子得分均较高的地区。读者可以自己根据因子得分聚四类并作散点图。

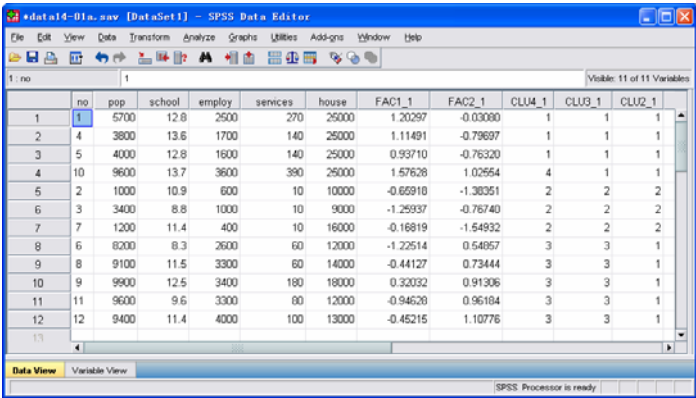


图 14-18 排序后的数据

通过以上分析可以看出，使用因子得分绘制出散点图是比较容易进行的，反过来对照原始数据，也可以得出同样的分析结论。但是直接使用 5 个原始变量来分类，就不够直观。由此我们对因子分析的作用也可以有较实际的体会了。

5. 排序后观察因子和原始数据

- (1) 按 Data→Sort Cases 操作，将 Clu3\_1 作为排序关键字，按升序排序。
- (2) 观察各类原始变量的特点也可以对各类特征得出明显的结论，见图 14-18。

应该说明的是，实际应用中，做因子分析要求观测量数至少应该是变量数的 5 倍以上。而本例 5 个变量 12 个观测量，所以仅作为一个介绍方法的例题而已。

14.1.5 市场研究中的顾客偏好分析

在市场研究中，常常要求分析顾客的偏好和当前市场的产品与顾客偏好之间的差别，从而找出新产品开发的方向。顾客偏好分析时常用到主成分分析方法。

下面的例题数据 data14-02 来自 SAS 公司。1980 年一个汽车制造商在竞争对手中选择了 17 种车型，访问了 25 个顾客，要求他们根据自己的偏好对 17 种车型打分。打分范围 0~9.9，9.9 表示最高程度的偏好。

- 1. 数据文件格式。数据文件是以 25 个顾客的评分分为 25 个变量，即 v1~v25，每

种车型的 25 个分数是一个观测量, 17 种车型为 17 个观测量。

## 2. 操作步骤

(1) 按 Analyze→Data Reduction→Factor 顺序单击菜单项, 打开因子分析主对话框。

(2) 选择 v1~v25 为分析变量送到右面的 Variables 栏中。

(3) 在主对话框中单击 Extraction 按钮, 相应的对话框中:

① Method 菜单中选择 Principle components 项, 使用主成分分析方法。

② Analyze 栏中选择 Correlation matrix 项, 分析相关矩阵。

③ Extract 栏中选择 Number of factors, 并输入 3。

④ Display 栏中选择 Unrotated factor solution, 显示未旋转的因子结果。同时选择 Scree Plot, 要求作出特征值的散点图。

⑤ Maximum iteration convergence 25, 结束迭代的判据为到达最大迭代次数 25。

(4) 主对话框中单击 Score 按钮。在相应的对话框中选择 Save as variables, 并在 Method 栏中选择 Regression, 要求通过回归方法计算因子得分并把因子得分作为变量保存到数据文件中。

(5) 单击 Descriptives 按钮, 在对话框中 Statistics 栏内不选择 Initial solution。

3. 主对话框中单击 OK 提交系统执行。输出窗中的语句如下:

```
FACTOR  /VARIABLES v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v14 v15 v16 v17 v18 v19 v20
        v21 v22 v23 v24 v25
        /ANALYSIS v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v14 v15 v16 v17 v18 v19 v20 v21
        v22 v23 v24 v25
        /PRINT EXTRACTION /PLOT EIGEN /CRITERIA FACTORS(3) ITERATE(25)
        /EXTRACTION PC /ROTATION NOROTATE /SAVE REG(ALL)/METHOD=CORRELATION .
```

4. 输出结果见表 14-16、表 14-17, 图 14-19、图 14-20。

## 5. 结果说明

选择提取公因子的方法为主成分法, 因此输出中的因子即成分。

表 14-16 是提取的三个因子的因子载荷矩阵。行列交叉点上的数据是对应因子在变量(顾客)上的载荷。它体现了交叉点对应的因子(列)与对应变量(行)的相关程度。

表 14-17 在选择提取公因子的数量时, 我们没有选择特征值大于 1 决定公因子数的方法, 而是选择了提取前 3 个公因子。此表为前 3 个因子所解释的原始变量的总方差, 及其占总方差的百分比和累计百分比。可以看出, 前 3 个因子(或成分)可以解释总方差的近 75%, 其余 22 个因子只占 25%, 可以说 3 个因子可以解释总方差的绝大部分。

图 14-19 是特征值碎石图。可以看出前 3 个特征值间的差异很大, 其余的变化很小。虽然也有特征值大于 1 的, 但变化量很小。从图中可以看出, 取前三个因子是正确的。

图 14-20 是当前数据文件。其中最右边的变量 Fac1\_1、Fac2\_1 和 Fac3\_1 是各观测量(17 种车型)的因子得分变量。该数据保存到 data14-02a 中。

根据数据文件中的因子得分变量和表 14-16 中的数据作散点图，得到偏好图。

表 14-16 初始因子载荷阵

	Component Matrix <sup>a</sup>		
	1	2	3
被访者1	.274	.625	.330
被访者2	.956	.068	-.210
被访者3	.778	-.300	-.151
被访者4	.491	.735	.343
被访者5	.451	.698	-.318
被访者6	.238	.677	-.059
被访者7	.783	-.212	.170
被访者8	.510	-.051	.713
被访者9	-.513	.718	-.189
被访者10	.936	-.191	.050
被访者11	.852	.143	-.260
被访者12	.836	-.085	-.356
被访者13	.943	.000	-.149
被访者14	.830	.198	-.081
被访者15	.858	-.174	-.067
被访者16	-.015	.803	.077
被访者17	.105	.658	.235
被访者18	.717	.609	.096
被访者19	.779	.126	-.033
被访者20	.773	-.570	.124
被访者21	.071	.657	-.095
被访者22	.238	-.459	.753
被访者23	-.766	.333	.281
被访者24	-.162	-.753	-.209
被访者25	-.765	.158	-.270

Extraction Method: Principal Component Analysis.  
a. 3 components extracted.

表 14-17 前三个因子（或成分）的方差解释

Component	Total Variance Explained		
	Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	10.837	43.348	43.348
2	5.802	23.207	66.555
3	2.060	8.240	74.795

Extraction Method: Principal Component Analysis.

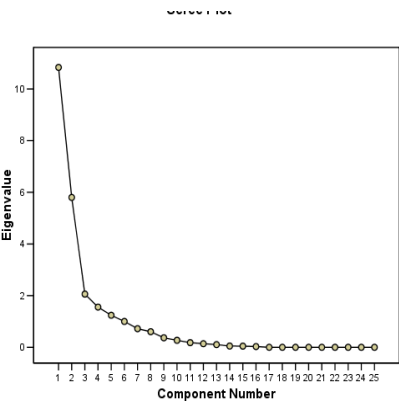


图 14-19 特征值散点图

6

6. 做偏好图

根据数据文件中的前两个因子得分变量作 17 个车型的散点图。也可先读取 data14-02a，步骤如下：

(1) 按 Graphic→Legacy Dialogs→Scatter/Dot...顺序单击菜单项，展开 Scatter/Dot...对话框。选择左上角的 Simple Scatter 项，单击 Define 按钮，打开 Simple Scatterplot 简单散点图对话框，如图 14-14 所示。

(2) 将 Fac1\_1 (REGR factor 1 for ...)送入 X Axis 作为 X 轴变量，

将 Fac2\_1 (REGR factor 2 for ...)送入 Y Axis 作为 Y 轴变量，变量 name 送入 Set markers by 栏中。单击 OK 按钮得到如图 14-21 所示的散点图。

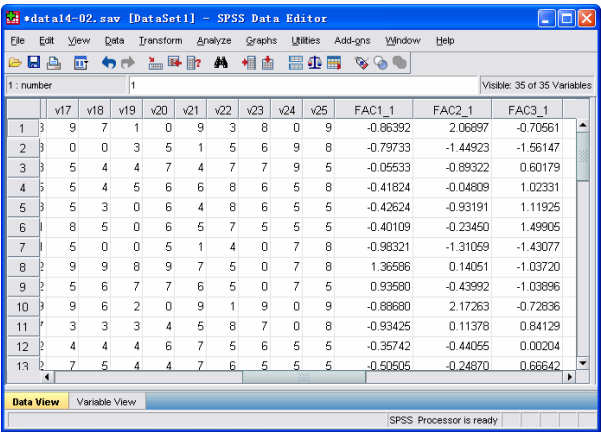


图 14-20 数据文件中的三个新变量：因子得分

将 Fac2\_1 (REGR factor 2 for ...)送入 Y Axis 作为 Y 轴变量，变量 name 送入 Set markers by 栏中。单击 OK 按钮得到如图 14-21 所示的散点图。

图中的字母均为变量 **name** 的值。

(3) 根据表 14-16 的因子载荷数据, 该数据见 data14-02b。用同样方法作另一个散点图, 见图 14-22。

### 7. 分析与结论

结合输出表, 比较图 14-21 和图 14-22。如果有条件, 可以将两张图的坐标原点对齐, 并经过透明处理, 则更便于比较。可以看出:

(1) 图 14-21 是根据 17 种车型的前两个因子得分作的图。

第一因子反映了车的产地。分数最高的是 DL 点 (沃尔沃), 最低的是 P 点 (福特)。横坐标右端多为欧洲车 (D、R) 两种大众车或日本车 (A、CI) 两种本田车, 在坐标左段多为美国车 (P) 福特、(CH) 雪伏龙等, 各自的第一因子得分说明顾客对欧洲车和日本车的评价较高。

第二因子反映了车的特性: 质量、动力、座位数等。分数高的是 (Co) 林肯、(E) 凯迪拉克, 位于坐标高端, 分数低的为 (P) 福特、(CH) 雪伏龙, 说明顾客对高档车的质量评价较高。

(2) 在图 14-22 中。与图 14-21 对应着进行综合分析, 图 14-22 的点在第二象限的 (左上方) 的顾客偏好大型豪华美国车; 点在第四象限的很多, 这些顾客偏好日本和欧洲车; 第三象限的点很少, 说明顾客中偏好美国小型车的很少。图 14-22 第一象限点很多, 但相应图 14-21 中的第一象限车很少。这可能预示着新车型产品市场或该汽车生产商的主要竞争对手没有相应的产品。这正是新产品开发的方向: 高质量、豪华大型欧洲、日本车。

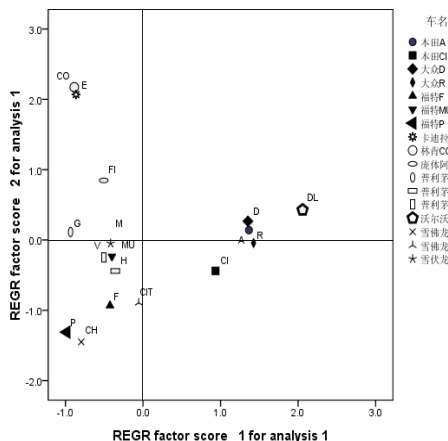


图 14-21 17 种车型的因子得分散点图

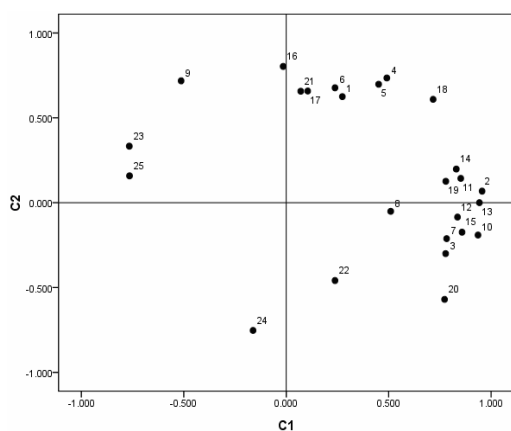


图 14-22 25 个顾客的偏好散点图

1980 年的此项研究虽然使用的样本很小, 但事后发现, 这项研究对市场的分析相当



准确: 目前除美国福特车外, 其他中小型车 (MU 点) 几乎都不生产了; 日本车在 20 世纪 80 年代的市场占有率是比较高的; 日本和欧洲汽车制造商开发了大型豪华车如德国的宝马、日本的凌志, 都是很受顾客欢迎的车型。

## 14.1.6 因子分析过程的命令语句

### 1. 因子分析过程命令语句

FACTOR VARIABLES=varlist?

```
[/MISSING={LISTWISE**}{PAIRWISE}{MEANSUB}{DEFAULT**}] [INCLUDE]]
[/MATRIX={IN ({COR=file}{COR=*}{COV=file}{COV=*}{FAC=file}{FAC=*})}
          {OUT({COR=file}{COR=*}{COV=file}{COV=*}{FAC=file}{FAC=*})}]
[/METHOD = CORRELATION**][COVARIANCE]]
[/SELECT=varname (value)] [/ANALYSIS=varlist...]
[/PRINT={DEFAULT**}[INITIAL**][EXTRACTION**][ROTATION**][UNIVARIATE]
        [CORRELATION][COVARIANCE][DET][INV][REPR][AIC][KMO] [FSCORE]
        [SIG] [ALL]]
[/PLOT={EIGEN} [ROTATION [(n1,n2)]]] [/DIAGONAL={value list} {DEFAULT**}]
[/FORMAT={SORT} [BLANK (n)] [DEFAULT**]]
[/CRITERIA={FACTORS (n)} [MINEIGEN ({1.0**}{n})] [ITERATE ({25**}{n})]
[RCONVERGE ({0.0001**}{n})] [{KAISER**}{NOKAISER}]
[ECONVERGE ({0.001**}{n})] [DEFAULT**]]
[/EXTRACTION={PC**}{PA1**}{PAF}{ALPHA}{IMAGE}{ULS}{GLS}{ML}
              {DEFAULT**}]
[/ROTATION={VARIMAX**}{EQUAMAX}{QUARTIMAX}{OBLIMIN
            ({0})}{n}{PROMAX ({4} {n})}{NOROTATE}{DEFAULT**}]
[/SAVE={REG}{BART}{AR}{DEFAULT} ({ALL} {n} [rootname])]]
```

其中“?”表明省略 VARIABLE 选项时使用的输入矩阵。“\*\*”表明当该子命令省略时的系统默认值。

“[ ]”中的均为子命令, 一个子命令中的若干个并列的 “[ ]”中的内容表明该子命令完成其功能需要指定的参数, 是子命令下属的第一层选项, 并列的几个选项可以取其一, 也可以取其中的若干项; 在这一层选项下面并列的若干个用 “{ }”包含的内容是第二层选项, 并列的第二层选项中只能取其一。

FACTOR 过程根据相关矩阵或协方差矩阵, 完成因子分析。提供七种提取因子的方法, 可以选择其中一种方法。FACTOR 过程还可以接受相关矩阵、协方差矩阵或因子载荷矩阵方式的输入矩阵, 也可以把矩阵资料写入一个矩阵数据文件。

因子分析使用 FACTOR 命令语句调用相应的过程, 命令语句格式如下:

**FACTOR VARIABLES=**变量表

**FACTOR** 命令语句指定参与因子分析的变量, 但当使用输入矩阵时这一部分可以省略。使用 **VARIABLES=**变量表的形式指定参与分析的变量时, 数据文件必须在数据窗口中。变量表中的变量必须是数据窗口中的变量, 而且必须是观测的变量。

## 2. 因子分析过程中的子命令

子命令按照清单给出的顺序出现不会出错。

**FACTOR** 的子命令很多, 使用的条件各不相同。下面按功能模块分别介绍。

(1) **MISSING** 子命令, 指定分析过程中处理缺失值的方法, 常用的处理方法有:

- **LISTWISE**, 是系统默认的处理方法。从分析中剔除所有在分析变量表列出的变量中带有缺失值的观测量。

- **PAIRWISE**, 指定仅剔除正在计算相关系数的两个变量中, 其值缺失的观测量。

- **MEANSUB**, 使用各变量的均值代替缺失值进行计算。

- **INCLUDE**, 分析时对带有读者定义的缺失值观测量不剔除。一般不采用此方法。

(2) **WIDTH** 子命令, 指定输出宽度, 可以有以下几种指定方式:

- 132, 指定输出宽度为 132 列。

- $n$ , 指定输出宽度为  $n$  列,  $n$  为正整数。

- **DEFAULT**, 系统默认的输出宽度是 80 列, 或者使用在 **Preference** 的 **Output** 选项中指定的输出页的宽度。

(3) **MATRIX** 子命令, 指定输入/输出的矩阵数据或数据文件。在关键字 **IN** 后面的圆括号中指定输入的矩阵数据或数据文件, 在关键字 **OUT** 后面的圆括号中指定输出的矩阵数据或数据文件。无论输入还是输出, 均有以下 3 种选择:

- **COR=file**, 在等号后面指定保存相关矩阵的数据文件名, 可以包括保存路径。

- **FAC=file**, 在等号后指定保存因子载荷矩阵的数据文件名, 可以包括保存路径。

- **COV=file**, 在等号后面指定保存协方差矩阵的数据文件名。

如果以上 3 个选项等号后面是 “\*”, 对于输入矩阵 (**IN**), 星号表示使用数据编辑窗口中的矩阵; 对于输出 (**out**) 矩阵, 星号表示结果矩阵文件输出到当前数据窗口中。

(4) **ANALYSIS** 子命令, 指定此次参与分析的变量表, 在等号后面列出变量名。一次调用 **FACTOR** 过程可以使用一个以上的 **ANALYSIS** 命令, 从而达到给定不同参数进行多次分析的目的。

(5) **PRINT** 子命令, 指定要求计算和输出的统计量。

**DEFAULT**, 系统默认的输出项, 一般包括 **INITIAL**、**EXTRACTION** 等的输出结果。

- **INITIAL**, 初始分析结果, 选择此项可以给出原始变量的公因子方差、与变量数目相等的因子、各因子的特征值、各因子特征值占总方差的百分比及累积百分比。

- **EXTRACTION**, 未加旋转的因子提取结果。

- **ROTATION**, 旋转结果。指定此项将对正交旋转显示旋转后的因子矩阵模式、因

子转换矩阵, 对斜交旋转显示旋转后的因子矩阵模式、因子结构矩阵和因子间的相关矩阵。

- **UNIVARIATE**, 要求输出单变量描述统计量, 选择此项可以输出参与分析的各原始变量的均值、标准差等。

- **CORRELATION**, 相关系数, 选择此项给出原始变量间的相关系数矩阵。

- **DET**, 相关系数矩阵的行列式。

- **INV**, 相关系数矩阵的逆矩阵。

- **REPR**, 再生相关矩阵, 选择此项给出因子分析后的相关矩阵, 还给出残差, 即原始相关与再生相关之间的差值。

- **AIC**, 反映像相关矩阵, 包括偏相关系数的负数; 反映像协方差矩阵, 包括偏协方差的负数。

- **KMO**, **KMO** 和球形 **Bartlett** 检验, 选择此项给出 **Kaisex-Meyer-Olkin**, 它是对采样充足度的测度。

- **FSCORE**, 因子得分。

- **SIG**, 显著性水平, 选择此项给出每个相关系数相对于相关系数为 0 的假设检验的概率水平。

- **ALL**, 所有以上选项指定的统计量。

(6) **PLOT** 子命令, 指定要求绘制的统计图, 可以指定绘制:

- ① **EIGEN**, 要求显示碎石图。该图的两个坐标轴, 一个是按特征值大小排列, 一个是按因子序号排列的。

- ② **ROTATION**, 要求显示进行因子旋转后的因子载荷散点图。指定此项将给出以两因子为坐标轴的各变量的载荷散点图。

- ③ **ROTATION(n1,n2)**, 给出原始变量在 **Factor-n1** 对 **Factor-n2** 坐标系中的散点图。

(7) **FORMAT** 子命令

**FORMAT** 子命令指定显示统计量的格式。最主要的选项是 **SORT**, 要求按因子载荷系数大小排列。**DEFAULT** 是系统默认的显示方式, 是按变量名的字符排列。

(8) **CRITERIA** 子命令, 指定控制因子分析过程结束的判据。可以选择控制因子分析进程或结果的方式和参数, 选项有以下六个:

- ① **FACTORS(n)**, 要求最后选取  $n$  个因子。

- ② **MINEIGEN(1.0)**或 **MINEIGEN(n)**, 指定提取因子时使用最小特征值为 1.0 的判据, 或者由读者自己指定特征值判据  $n$ 。

- ③ **ITERATE(25)**或 **ITERATE(n1)**, 指定使用迭代次数作为结束迭代过程的参数。系统默认迭代 25 次结束提取因子的迭代过程, 读者也可以自己指定迭代次数  $n1$ 。 $n1$  为正整数, 写在选项关键字后面的括号中。

- ④ **KAISER** 或 **NOKAISER** 指定使用或不使用凯泽准则。

⑤ ECONVERGE(0.0001)或 ECONVERGE( $n$ ), 指定用最大公因子方差的变化值作为因子分析的收敛判据, 系统默认值为 0.0001。读者指定的收敛判据放在关键字后面的括号中 ( $n$ )。

⑥ RCONVERGE(0.0001)或 RCONVERGE( $n$ ), 指定用最大公因子方差的变化值作为因子旋转的收敛判据。系统默认最大变化值小于 0.0001 时, 迭代停止。该值也可以由读者自己指定 ( $n$ )。

⑦ DELTA, 指定斜交旋转的 $\delta$ 值, 该值应该在 $-1\sim 0$ 之间。

(9) EXTRACTION 子命令, 指定提取因子的方法, 可以在以下 5 种方法中进行选择。

① PC, 主成分法, 是系统默认的提取方法。

② PAF, 主轴因子提取法, 使用多元相关的平方作为对公因子方差的初始估计。

③ ALPHA、IMAGE 分别为 $\alpha$ 因子提取法、映像因子提取法。

④ USL、ML 分别为不加权最小平方法、最大似然法。

⑤ GSL, 通用最小平方法, 是用变量的单值加权的综合最小平方法。

(10) ROTATION 子命令, 指定因子旋转方法, 可以选择:

① NOROTATE 不进行旋转

② VARIMAX、EQUAMAX 分别为方差最大旋转、平均正交旋转

③ QUARTIMAX 四次方最大正交旋转

④ OBLIMIN, 斜交旋转, 指定此项可以在 CRITERIA 子命令中指定 $\delta$ 值。

(11) SAVE 子命令该子命令指定:

① 哪些统计量要作为新变量保存到数据文件中。

• REG, 回归法计算出的因子得分。

• BART, 巴特利特法计算出的因子得分。

• AR, 安德森-鲁宾法计算出的因子得分。

• DEFAULT, 系统默认的计算方法, 一般为回归法。

② 指定保存因子得分变量数目, 放在方法关键字后面的括号中。

• ALL, 指定保存所有主因子的得分。

•  $n$ , 使用一个正整数表明保存前 $n$ 个因子的得分。

③ rootname 如果不想使用系统默认变量命名因子得分变量, 可以指定因子得分变量名的根名。系统自动在根名后面加序号作为变量名的一部分。

以上介绍的子命令中, ANALYSIS、CRITERIA、EXTRACTION、ROTATION、SAVE 子命令可以在程序中重复使用。在每个 ANALYSIS 子命令中指定一个分析和分析中使用的变量, 随后即可用 CRITERIA 指定此次分析的判据。使用 EXTRACTION 子命令指定此次分析提取因子的方法, 使用 ROTATION 子命令指定此次分析中使用的旋转方法, 使用 SAVE 子命令指定此次分析要保存的新变量。

## 14.2 对应分析

### 14.2.1 对应分析概述

#### 1. 对应分析的思路

对应分析也称为相应分析，是在 R 型和 Q 型因子分析的基础上发展起来的一种多元统计方法。它首先由法国统计学家 J. P. Beozecri 于 1970 年提出。

因子分析根据研究对象的不同而分为研究指标（变量）的 R 型因子分析和研究样品的 Q 型因子分析，使用因子分析方法时这两个过程只能分开进行。这样，一方面会漏掉一些指标（变量）和样品间的信息，另一方面因因子分析要求观测（样品）数目必须是变量数的 5 倍，因此，在做 Q 型因子分析时，还要做比 R 型因子分析计算量更大的计算工作。因此从研究设计的要求来说，并不是最佳的，所以，有必要改良算法，从而达到总计算量最小而又同时考虑指标（变量）和样品的关系。

对应分析借助列联表独立性检验中卡方统计量的计算方法，对原始数据矩阵进行转换，公式是

$$p_{ij} = x_{ij} / \sum_i \sum_j x_{ij}$$

由此得到一个规格化的“概率”矩阵，使数据资料具有对称性，当数据资料具有对称性时，量纲的差异也被消除，R 型和 Q 型因子分析之间就建立起了联系，在做 R 型因子分析时也就同时完成了 Q 型因子分析的工作，克服了由于样品容量大带来的 Q 型因子分析计算量大的困难。

另外，根据 R 型因子分析和 Q 型因子分析的内在联系，可在同一个坐标轴图形中将指标（变量）和样品同时反映出来，图形中邻近的变量点表示它们关系密切可分为一类，同样，邻近的样品点表示它们关系密切可归为一类，而且属于同一类型的样品点可用邻近的变量点来表征。

对应分析的目的之一是在同时描述各个变量类别之间的关系时，在一个低维度空间中对对应表里的两个名义变量之间的关系进行描述。对每个变量而言，图中类别点之间的距离反映邻近有相似分类图的各类别之间的关系。一个变量在从原点到另一个变量分类点的向量上的投影点描述了变量之间的关系。

很多学者认为对应分析方法是探索性数据分析的内容，因此，极大部分的使用者只要能够理解对应分析行、列记分图所包含的信息即可。

#### 2. 对应分析中需要考虑的事项

(1) 数据。用于分析的分类变量是名义尺度。对合计数据或对除频数以外的对应测度（correspondence measure），使用正相似性值的权重变量。

(2) 有关程序。如果包含的变量超过两个，使用多重（多元）对应分析。如果是有

序尺度变量，则使用分类主成分分析（Categorical Principal Components Analysis）。

## 14.2.2 对应分析过程

### 1. 对应分析数据预处理

在对应分析中，原始数据必须整理成交叉表的单元格计数形式。在对应分析前先用 **WEIGHT** 命令来进行处理。因此，在对应分析的数据文件中，需定义三个变量。将要放在对应分析过程中的行和列的变量是分类变量。第三个变量是对应行、列的实际测试值，一般为尺度变量。

在进行对应分析前，应先用 **Data** 菜单中 **Weight Cases** 功能定义权重变量。方法见 2.4.1 节。如果权重变量中有 0 值，会发出警告，但不影响对应分析的正常分析工作。

### 2. 操作步骤

(1) 按 **Analyze**→**Data Reduction**→**Correspondence Analysis** 顺序打开如图 14-23 所示的对应分析主对话框。

(2) 从变量表中选择行、列变量，送入 **Row** 和 **Column** 框中。

(3) 按 **Define Ranges** 按钮，见图 14-24，可定义行（列）变量参与分析的分类范围。在 **Minimum value** 中输入分类的最小值；在 **Maximum value** 中输入最大值。这两个值必须是整数。否则在分析中会删除小数部分。单击 **Update** 按钮，可将定义的分类数据上传到 **Category Constraints** 的框中。在分析中忽略在指定范围以外的分类值。

(4) **Category Constraints** 栏，定义类别的等同约束。所有类别最初没有约束。你可以约束某个行（列）类别去等于其他的行（列）类别，或者你可以定义一个行（列）类别作为辅助行（列）类别。如果分类值所代表的类别不符合分析需要或者界限是模糊的，可以使用等同约束将这样的类视为等同，即有相等记分的类。它共有三个选项：

① **None**，默认项，即分类数据保持原状，不作任何约束。

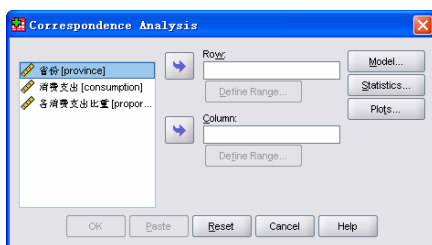


图 14-23 对应分析的主对话框

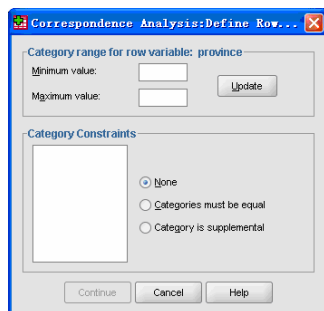


图 14-24 定义分类变量的范围

② **Categories must be equal**，类别必须有相等的记分。假如类别的次序是不想要的或违反直觉的，则可以使用等同约束。这可从 **Update** 产生的分类值列表中选择类别，指定等同约束，至少有两个类别必须是相等的。能用等同约束的行（列）分类的最大数量

是有效行(列)类别总数减 1。在分类集中为强加不同的等同约束,使用语句。例如,使用语句约束分类 1 等于分类 2 和分类 3 等于分类 4 (在 Syntax 编辑窗口中,加语句/EQUAL=变量名(1 2),(3 4))。

③ Category is supplemental,从 Update 产生的分类值列表中选择类别指定辅助类别。辅助类别不影响分析,只在由有效分类定义的空间里被描述。辅助类别在定义的维度数里不扮演角色。最大辅助的行(列)类别的数量是行(列)类别总数减 2。

(5) 指定对应分析模型。单击 Model 按钮进入 Model 对话框,见图 14-25。允许指定维度数、距离测度、标准化方法和正规化方法。

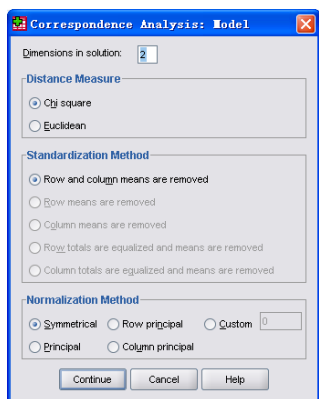


图 14-25 模型选项对话框

① Dimensions in solution 栏,指定对应分析解的维度数,默认值为 2。通常,选择为解释大多数变差所需要的较少的维度数。最大维度数取决于用于分析的有效分类数和等同约束数。最大维度数是下列中较小的一个:

- 有效的行分类数减去被等同约束的行分类数加约束的行分类集数;
- 有效的列分类数减去被等同约束的列分类数加约束的列分类集数。

② Distance Measure 栏,选择对应表的行间距离和列间的距离测度。

- Chi square 卡方。距离测度用加权距离,这里的权就是行或列的质量(边际概率)。标准对应分析要求该测

度。系统默认本方法。

- Euclidean 欧氏距离。两行之间或两列之间的差的平方和的平方根作为距离测度。

③ Standardization Method 栏,标准化方法选项。可从下列五个选项中选择一种:

- Row and column means are removed,行和列两者被中心化。标准对应分析需用本方法。当选用 Chi square 作为 Distance Measure 的选项时,系统只默认本方法。

- Row means are removed,只有行被中心化。
- Column means are removed,只有列被中心化。
- Row totals are equalized and means are removed,先使行边际相等,再中心化行。
- Column totals are equalized and means are removed,先使列边际相等,再中心化列。

④ Normalization Method,正规化方法选项。可从下列五个选项中选择一种:

- Symmetrical 对称法。对各个维度,行记分是列记分除以匹配奇异值的加权平均,列记分是行记分除以匹配奇异值的加权平均。使用本方法可以检查两个变量分类间的差异或相似。

- Principal,行点和列点之间的距离是与选定的距离测度一致的对应表中距离的近似

值。如果要检查一个或两个变量的类别之间的差异，而不是两个变量之间的差异，使用本方法。

- **Row principal**，行分数间的距离是在对应表中根据选定方法对距离测度的近似值。行记分是列记分的加权平均。要检查行变量的类间差异或类似程度，使用本方法。

- **Column principal**，列分数间的距离是在对应表中根据选定方法计算的距离的近似值。列记分是行记分的加权平均。要检查列变量的类间差异或类似程度，使用本方法。

- **Custom** 自定义。必须在  $\pm 1$  间指定一个值。值  $-1$  对应于主要列。值  $1$  对应于主要行。值  $0$  对应于对称的。所有其他的值传达行和列分数变化程度的惯量。本方法通常用来制作特制的二维图形。

(6) 按 **Statistics** 选项按钮，进入 **Statistics** 对话框，见图 14-26。指定输出哪些结果表。

① **Correspondence table**，要求输出含有变量行和列边际总和的交叉分组列表。

② **Overview of row points**，要求输出行综合表，表中包括行变量各分类的记分、质量、惯量、分数对维度惯量的贡献、维度对分数惯量的贡献。

③ **Overview of column points** 列分数综述选项。在输出窗口中为各个列分类显示包括记分、质量、惯量、分数对维度惯量的贡献、维度对分数惯量的贡献的综合表。两个输出选项：

- **Row profiles**，行归一化处理后的分布表。
- **Column profiles**，列归一化处理后的分布表。

④ **Permutations of the correspondence table**，输出按第一维度上记分的递增顺序排列的行、列对应表。在任选项中，可为将要产生的序列改变的表指定最大维度数。为各维度产生一个从 1 到指定数目的序列改变表。

⑤ **Confidence Statistics for**，在本选择中共有两个选项：

- **Row points**，输出包括标准差和所有非辅助行分数相关内容的表格。
- **Column points**，输出表格包括标准差和所有非辅助列分数相关内容。

(7) 统计图选项。按 **Plots** 选项按钮，进入 **Plots** 对话框，见图 14-27。

① **Scatterplots** 栏，散点图选项产生矩阵的所有维度的成双图。共有三个选项：

- **Biplot**，双维图法。输出矩阵的行、列分数联合图。如果选择了 **Principal** 正规化方法，则本选项无效。

- **Row points**，输出矩阵的行分数的图。
- **Column points**，输出矩阵的列分数的图。

- **ID label width for scatterplots**，设置散点图中 ID 标签宽度，默认值 20。该值必须是小于等于 20 的正整数。

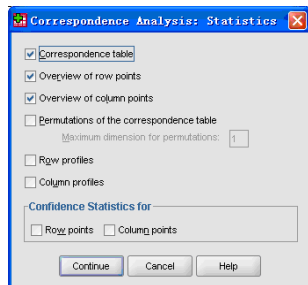


图 14-26 选项对话框



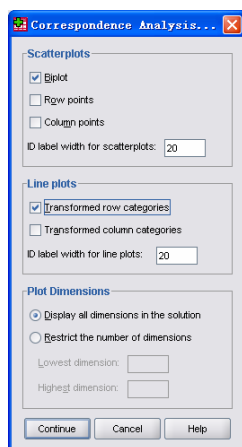


图 14-27 Plots 对话框

② Line plots 栏, 产生所选变量每一个维度的线图。有以下两个线图供选择。

- Transformed row categories, 输出行分类转换图。行分类值取决于相应的行记号。

- Transformed column categories, 输出列分类转换图。列分类值取决于相应列记号。

- ID label width for scatterplots 参数框, 指定线图 ID 标签宽度, 默认值 20。指定值必须是小于等于 20 的正整数。

③ Plot Dimensions 栏, 允许去控制在输出中显示的图的维度。有两个选项:

- Display all dimensions in the solution, 在散点图矩阵里显示解中的所有维度。

- Restrict the number of dimensions 复选项, 显示的维度被限制在成对图。如果限制维度, 则必须选择作图的最低和最高维度。最低维度可从 1 到解的维度数减 1 的范围中取值, 并且所作的图以较高维度为背景。最高维度值可从 2 到解的维度数的范围中取值, 并指出被使用于成对维度图中的最高维度。本说明适用于所有要求的多维图。

### 14.2.3 对应分析实例

【例 4】用对应分析的方法研究我国部分省份的农村居民人均消费支出结构。数据资料来源于《中国统计年鉴》1997 年。

在 data14-03 数据文件中, 共有三个变量, 分别为: province (省份: 1 山西、2 内蒙古、3 辽宁、4 吉林、5 黑龙江、6 海南、7 四川、8 贵州、9 甘肃、10 青海)、consumption (消费支出分类: 1 食品、2 衣着、3 居住、4 家庭设备及服务、5 医疗保健、6 交通和通讯、7 文教娱乐)。Proportion (各种消费支出比重) 是尺度变量。前两个为名义变量。

首先用 data→weight cases 功能定义 proportion 各种消费支出比重变量为权重变量。

1. 按 Analyze→Data Reduction→Correspondence Analysis 顺序单击菜单项, 进入对应分析的主对话框。

2. 选择 province 省份变量, 移入 Row 的下框中, 按其下的 Define Ranges 按钮, 在 Minimum value 中输入 1, 在 Maximum value 中输入 10; 按 Update 按钮, 送到 Category Constraints 框中。由于没有辅助项、等约束项及强制性等约束项, 因此, 在 Category Constraints 的选项中使用系统默认选项 None。

3. 选择 consumption, 移入 Column 框中, 按 Define Ranges 按钮, 在 Minimum value 中输入 1, 在 Maximum value 中输入 7, 按 Update 按钮, 送到 Category Constraints 框中。

4. 按 Model 按钮选项, 进入 Model 对话框。

• Dimensions in solution 参数框, 由于本例中样品数和指标(变量)数都较少, 使用系统默认值 2, 即将样品和变量对应地分为两类。

• 在 Distance Measure 栏, 选择系统默认的 Chi square 卡方距离测度。

• 在 Standardization Method 栏, 由于在上面的距离测度选项中选定了卡方距离测度, 故在此只能选择系统默认的 Row and column means are removed 项。

• 在 Normalization Method 栏, 选择 Symmetrical 项。

5. 按 Statistics 按钮, 在 Statistics 对话框中, 选择 Correspondence table, 要求输出对应表、选择 Row profiles、Column profiles, 输出行、列变量归一化处理表。

6. 按 Plots 按钮, 进入 Plots 对话框, 只选择 Biplot 复选项。

7. 单击 Paste 按钮, 可生成命令语句。按 OK 按钮, 执行运算。

8. 输出结果, 见表 14-18 至表 14-21 和图 14-28。

结果解释:

输出结果包括反映原始数据组成的对应表、行和列的归一化处理表及汇总表。

表 14-18 对应表给出 10 个省份的 7 种消费支出的观察值、总和, 行、列有效边际值 Active Margin。最右下角的值 9.825 是所有观察值的和。

表 14-18 对应表

Correspondence Table								
	消费支出							
省份	食品	衣着	居住	家庭设备及 服务消费	医疗 保健	交通 通讯	文教 娱乐	Active Margin
山西	.584	.111	.092	.050	.038	.019	.080	.975
内蒙	.581	.081	.112	.042	.043	.040	.083	.984
辽宁	.565	.100	.124	.041	.043	.031	.079	.984
吉林	.531	.105	.117	.045	.044	.039	.095	.976
黑龙江	.555	.097	.143	.038	.052	.026	.073	.984
海南	.655	.048	.095	.048	.022	.019	.087	.983
四川	.640	.062	.117	.048	.034	.017	.072	.990
贵州	.725	.056	.073	.044	.016	.016	.057	.989
甘肃	.679	.050	.088	.038	.040	.015	.068	.978
青海	.666	.089	.097	.038	.039	.019	.034	.982
Active Margin	6.181	.800	1.060	.433	.372	.241	.738	9.825

表 14-20 列归一化处理表

Column Profiles								
	消费支出							
省份	食品	衣着	居住	家庭设备及 服务消费	医疗 保健	交通 通讯	文教 娱乐	Mass
山西	.094	.139	.087	.116	.103	.078	.108	.099
内蒙	.094	.102	.106	.098	.116	.166	.113	.100
辽宁	.091	.125	.117	.095	.117	.130	.107	.100
吉林	.086	.132	.110	.104	.118	.160	.129	.099
黑龙江	.090	.121	.135	.087	.140	.109	.099	.100
海南	.106	.060	.090	.111	.060	.077	.131	.100
四川	.104	.077	.110	.112	.090	.072	.098	.101
贵州	.117	.070	.069	.102	.044	.065	.078	.101
甘肃	.110	.063	.083	.088	.107	.063	.092	.100
青海	.108	.111	.091	.088	.106	.080	.046	.100
Active Margin	1.000	1.000	1.000	1.000	1.000	1.000	1.000	

表 14-19 行归一化处理表

Row Profiles								
	消费支出							
省份	食品	衣着	居住	家庭设备及 服务消费	医疗 保健	交通 通讯	文教 娱乐	Active Margin
山西	.599	.114	.095	.051	.039	.019	.082	1.000
内蒙	.591	.083	.114	.043	.044	.041	.085	1.000
辽宁	.574	.102	.126	.042	.044	.032	.080	1.000
吉林	.544	.108	.120	.046	.045	.039	.098	1.000
黑龙江	.564	.098	.146	.038	.053	.027	.074	1.000
海南	.666	.049	.097	.049	.023	.019	.098	1.000
四川	.647	.062	.118	.049	.034	.018	.073	1.000
贵州	.734	.057	.074	.045	.017	.016	.058	1.000
甘肃	.694	.052	.090	.039	.041	.016	.069	1.000
青海	.678	.090	.099	.039	.040	.020	.034	1.000
Mass	.629	.081	.108	.044	.038	.025	.075	

表 14-21 汇总表

Summary								
Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
1	.133	.018			.657	.657	.309	
2	.070	.005			.180	.837	.294	
3	.050	.002			.091	.928		
4	.036	.001			.047	.976		
5	.024	.001			.021	.996		
6	.010	.000			.004	1.000		
Total		.027	.265	1.000 <sup>a</sup>	1.000	1.000		-.050

a. 54 degrees of freedom

表 14-19 为对应表中每行观察值除以每行总和的归一化结果。行的边际都为 1。

表 14-20 为对应表中每列观察值除以每列总和的归一化结果。每列的边际都为 1。

表 14-21 为汇总表, 给出了行与列记分之间的关系, 从左到右各列依次为维度、奇

异值（即惯量的平方根，反映了行与列各水平在二维图中分量的相关程度，是行与列进行因子分析产生新的综合变量的典型相关系数）、惯量（为每一维到其重心的加权距离的平方，它用来度量行列关系的强度）、卡方值（即列联表行列独立性卡方检验的卡方值）、显著性水平（即行列独立的零假设下的概率值，值很大说明列联表的行与列之间独立，否则有较强的相关性）、惯量比例（是各维度即公因子分别解释总惯量的比例及累计百分比，类似因子分析中公因子解释能力的说明）。

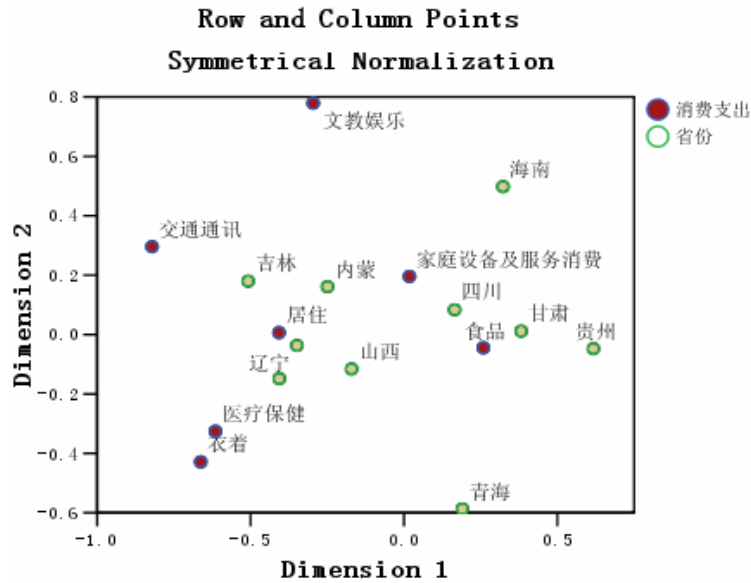


图 14-28 行和列记分图

从该表中可见，由于第一维（0.657）、第二维（0.180）的惯量比例和 83.7%，因此其他维度的重要性可以忽略。

- 在图 14-28 中，以横轴 0 为中心轴，可将变量点和样品点分为两类。
- 第一类：变量为衣着、居住、医疗保健；省份有山西、内蒙古、辽宁、吉林、黑龙江，它们位于我国的东部和北部地区，说明这 5 个省份的消费支出结构相似。
- 第二类：变量为食品、家庭设备及服务支出比重；省份有四川、贵州、甘肃。它们位于我国西部和南部地区，说明这 3 个省份的消费支出结构相似。
- 青海、海南距各种消费类型都较远，较特殊，但这两个省份距离较远，类型又不同。

14.2.4 对应分析过程的命令语句

1. 对应分析过程的命令语句清单如下：

```
CORRESPONDENCE/TABLE={row var (min, max) BY column var (min, max)}{ALL (# rows, # columns)}  
[/SUPPLEMENTARY=[{row var (valuelist)} {ROW (indexlist)}] [{column var (valuelist)}]
```

```

{COLUMN (indexlist)}}]
[/EQUAL=[{row var(valuelist)(valuelist)}{ROW(indexlist) (indexlist)}]
[{Column var (valuelist)...(valuelist)}{COLUMN(indexlist) (indexlist)}]
[/MEASURE={CHISQ**} {EUCLID}]/[DIMENSION={2**}{value} ]
[/STANDARDIZE={RMEAN}{CMEAN}{RCMEAN**}{RSUM}{CSUM}]
[/NORMALIZATION={SYMMETRICAL**}{PRINCIPAL}{RPRINCIPAL} CPRINCIPAL}{value}]
[/PRINT=[TABLE**] [RPROF] [CPROF] [RPOINTS**] [CPOINTS**] [RCONF] [CCONF]
[PERMUTATION (1***)] [DEFAULT] [NONE] ]
[/PLOT=[NDIM=({1**, 2**}{value, value}{value, MAX})] [RPOINTS [(n)]
[CPOINTS [(n)]] [TRROWS [(n)]] [TRCOLUMNS [(n)]] [BIPLOT**[(n)]] [NONE]]
[/OUTFILE={SCORE (filename)} {VARIANCE (filename)} {SCORE (filename)
VARIANCE (filename)}}].

```

标有“\*\*”的是默认值，子命令或关键字被省略时执行。

## 2. CORRESPONDENCE命令语句功能如下：

CORRESPONDENCE利用散点图矩阵表示一个二维表格的行和列间的关联。它计算行和列的得分、统计和根据得分制作图形及计算置信统计量。它和TABLE子命令是基本语句。在默认情况下，它计算一个二维解和显示对应分析表、汇总表、行和列分数的概述和关于二维中的第一个维度上的行和列得分的二维散点矩阵图。TABLE子命令必须首先出现。其他的子命令出现顺序任意。MEASURE、STANDARDIZE和NORMALIZATION子命令中只能出现一次。如果超过一个，执行最后出现的子命令。CORRESPONDENCE将把表格数据和合计数据负数值作为0处理。指定作为增补的行和列不能相等。变量增补最大数量是200，变量等同的最大数量是200。

## 3. 子命令

(1) TABLE子命令指定行和列变量和它们的整数值范围。两个变量由关键字BY分开。每个变量后括弧中是指定参与分析的数值范围：最小值和最大值。

表格数据可以使用TABLE后的关键字ALL直接读取和分析。详见语句帮助文件。

(2) DIMENSION 子命令后等号连接一个正整数，是要求计算的维度数。系统默认值为2。通常，应该选择需要解释最大变化的维度数。维度的最小数可以为1。最大数可以指定等于实际的行数和列数中的最小值减1。实际的行和列是分析中使用的非增补的行和列。如果指定的维数超过维度允许的最大值，则自动缩减为允许的最大值。

(3) SUPPLEMENTARY 子命令用来指定要作为增补对待的行和列。样品有效数据，用变量名，后接圆括号中的一个值列表表示。值必须在 TABLE 子命令中的行和列变量指定的值范围内；如果是表格数据，在本子命令中用 ROW 和/或 COLUMN，后接圆括号中的一个值列表表示。值列表描绘了表格输入数据的行或列索引。行或列增补的最大数量是实际的行或列的数量减2。增补的行和列不能相等。

(4) **EQUAL** 子命令指定有相等记分约束的行或列。在 **EQUAL** 后用变量名,后面圆括号中至少有两个值的一个列表说明样品有效值数据。值必须在 **TABLE** 子命令中为行或列变量指定的值范围内;如果是表格数据,在 **EQUAL** 用 **ROW** 和/或 **COLUMN**,后接圆括号中的一个值列表说明。值描绘了表格输入数据的行或列索引。不能增补要求有相同得分限制的行或列。行或列相等的最大数量是实际的行或列的数量减 1。

#### (5) **MEASURE** 子命令

该子命令指定行和列图表间的距离测度。可选择的关键字:

① **CHISQ**, 卡方距离。是加权距离,权重为行或列的质量。**CHISQ** 为默认选项。

② **EUCLID**, 欧氏距离。距离是两行或两列值之间差异的平方和的平方根。

(6) **STANDARDIZ** 子命令。该子命令指定标准化方法。如果 **MEASURE** 是 **CHISQ**, 标准化方法被自动设置为 **RCMEAN**, 对应于标准对应分析。

(7) **NORMALIZATION** 子命令对行、列记分进行正规化处理,只影响记分和置信统计,贡献和 **profiles** 不改变。可从 **SYMMETRICAL**、**PRINCIPAL**、**RPRINCIPAL**、**CPRINCIPAL** 四个选项中选择一项。第五种方法是用代码表示前面的方法。1 与 **RPRINCIPAL** 法等价,0 等价于 **SYMMETRICAL** 方法, -1 等价于 **CPRINCIPAL** 方法。通过在 -1 和 +1 间指定一个值,用户可以伸展行和列记分上的惯性去改变程度。本方法对精心制作的双维图是有用的。

(8) **PRINT** 子命令控制对应分析中计算的几个统计表的显示。总是要产生汇总表(奇异值、惯性、惯性总计的比例、惯性总计的累计比例、维度的最大数量的置信统计)。不指定 **PRINT**, 则显示输入表格、汇总表格、行分数概要表和列分数概要表。

以下的关键字是可用的: **TABLE**、**RPROF**、**CPROF**、**RPOINTS**、**CPOINTS**、**RCONF**、**CCONF**、**PERMUTATION(1)**、**DEFAULT**、**NONE**。

(9) **PLOT** 子命令指定要输出的得分图:行、列、行和列得分图、行和列的得分转换图。不指定 **PLOT**, 或没有选择关键字, 则只产生一个双维图。可用的关键字是: **NDIM**、**RPOINTS[(n)]**、**CPOINTS[(n)]**、**TRROWS[(n)]**、**TRCOLUMNS[(n)]**、**BIPLOT[(n)]**、**NONE**。

所有的关键字可以在后面圆括弧中给出 0~20 间的整数,作为图中值标签的字符数。空格作为字符计数。0 表示用值不用值标签。没定义值标签,则使用实际值。标签长度根据长度参数缩减。**TRROWS**和**TRCOLUMNS**产生线性图形。**RPOINTS**和**CPOINTS**产生散点图矩阵。**BIPLOT**产生一个双维图矩阵。对线性图形,值标签用作分类轴标签。对散点图矩阵和双维图矩阵,值标签用作图形中分数的标签。

另外可以用 **NDIM** 关键字指定图形的维数,默认 **NDIM(1,2)**。否则,第一个值必须是从 1 到维度数减 1 中的任何整数。第二个值必须是从 2 到维度数中的任何整数。第二个值必须大于第一个值。作为选择,关键字 **MAX** 可以用来代替维度解答中指定的最高值。对 **TRROWS**和**TRCOLUMNS**, 第一和第二个值指出作图的维度范围。对 **RPOINTS**、**CPOINTS**和**BIPLOT**, 第一和第二个值指出包含在散点图矩阵或双维图里的维度范围。

(10) OUTFILE 子命令可向矩阵数据文件中写计算结果。在括号中给出外部文件名。

① OUTFILE 后面必须尾接以下两个关键词的一个或两个:

- SCORE(filename), 向矩阵数据文件中写行和列记分。
- VARIANCE(filename), 向矩阵数据文件中写方差和协方差。

如果同时指定SCORE和VARIANCE关键字, 两个文件名应有所区别。对VARIANCE而言, 在矩阵数据文件中不产生辅助和等同约束的行和列。

② 在 SCORE 矩阵数据文件中的变量有: ROWTYPE 字符串变量, 包括所有的行 ROW 和所有的列 COLUMN 的值。LEVEL 字符串变量, 包括各个原始变量的值或值标签。VARNAME 字符串变量, 包括原始变量名。DIM1...DIM $n$  包括各个维度行和列记分的数字型变量。各个变量被标识为 DIM $n$ , 其中  $n$  代表维度数。

③ 在VARIANCE矩阵数据文件里的变量和它们的值为: ROWTYPE\_字符串变量, 包含文件里所有样品的值COV; SCORE\_字符串变量, 包含值SINGULAR、行变量名或标签及列变量名或标签; LEVEL\_字符串变量包含行变量的值或标签、列变量的值或标签和score\_=SINGULAR的空值。

④ VARNAME\_字符串变量包含维度数。

⑤ DIM1...DIM $n$  包括各个维度方差和协方差的数字型变量。各个变量被命名为 DIM $n$ , 其中 $n$ 代表维度数。

## 习 题 14

1. 简述主成分分析的基本思想。
2. 用什么统计量衡量主成分中各成分提供的信息量?
3. 一般根据什么确定主成分提取的数量?
4. 简述因子分析的基本思想。
5. 为什么要对初始因子分析结果进行旋转?
6. 简述对应分析的基本思想, 对应分析与因子分析有什么不同?
7. 数据 data14-04 是某医院 3 年中各月的数据, 包括门诊人次、出院人数、病床利用率和周转次数、平均住院天数、治愈或好转率、病死率、诊断符合率、抢救成功率。采用因子分析法探讨综合评价指标。
8. 数据 data14-05 是 1997 年全国 31 个省市自治区按各种经济类型资产占总资产比重(%)的数据, 试对其作对应分析。

# 第 15 章 尺度分析

尺度分析在 Analyze 主菜单的 Scale 命令项中,其主要功能是进行信度分析和多维尺度分析。按 Analyze→Scale 顺序打开如图 15-1 所示的 Scale 菜单。Scale 包括的统计功能有四项:

1. Reliability Analysis 信度分析。
2. Multidimensional Unfolding 多维展示分析,该方法试图找到一种定量测度方法从而可以直观地研究两个事物之间的关系。
3. Multidimensional Scaling (PROXSCAL) 近似多维尺度分析。可以做相似性数据和不相相似性数据的分析,比 ALSCAL 功能更强。
4. Multidimensional Scaling (ALSCAL) 多维尺度分析。仅能分析不相相似性数据。

本章只介绍信度分析和多维尺度分析。

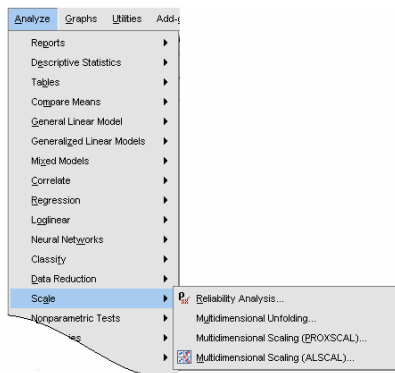


图 15-1 尺度菜单及其统计命令

## 15.1 信度分析

### 15.1.1 信度分析的概念

#### 1. 什么是信度

信度又叫可靠性,是指测验的可信程度。它主要表现测验结果的一贯性、一致性、再现性和稳定性。一个好的测量工具,对同一事物反复多次测量,其结果应该始终保持不变才可信。比如,我们用一把尺子测量一批物品,如果今天测量的结果与明天测量的结果不同,那么我们会对这把尺子的可信性产生怀疑。信度分析一般在心理学中应用较多,另外在学生考试试卷、社会问卷调查的有效性分析中也会涉及。信度只受随机误差影响,随机误差越大,测验的信度越低。因此,信度也可视为测量结果受随机误差影响的程度。系统误差产生恒定效应,不影响信度。

在测量学中,信度被定义为:一组测量分数的真变异数与总变异数(实得变异数)的比率,即

$$r_{xx} = \frac{S_r^2}{S_x^2}$$

式中的  $r_{xx}$  称作信度系数,  $S_r$  为真变异数,  $S_x$  为总变异数。

在实际测量中, 因为真值是未知的, 故信度系数不能由以上公式直接求出, 而只能根据一组实得分数(测得值)做出估计。

信度系数是衡量测验好坏的一个重要技术指标, 测验的信度系数达到多高才可以接受呢? 最理想的情况是  $r=1$ , 但这是办不到的。大多数学者认为: 任何测验或量表的信度系数如果在 0.9 以上, 则该测验或量表的信度甚佳; 信度系数在 0.8 以上都是可以接受的; 如果在 0.7 以上, 则该量表应进行较大修订, 但仍不失其价值; 如果低于 0.7, 量表就需要重新设计了。在心理学中通常可以用已有的同类测验作为比较的标准。一般能力与成就测验的信度系数常在 0.90 以上, 性格、兴趣、态度等人格测验的信度系数通常在 0.80~0.85 之间。

## 2. 相关术语

(1) 量表(scale): 用以测量的准尺。它是一个具有单位和参照点的连续体, 将被测量的事物置于该连续体的适当位置, 看它离开参照点多少单位的计数, 便得到一个测得值。这种连续体就称为量表。一般量表都由一套测验题目构成, 其中每一测验题都符合标准化要求, 具有一定的分值。

(2) 平行测量: 在心理学中能以相同的程度测量同一心理特质的测验。简单地说就是两个或两个以上的等值测量。同一特质的两个测量, 若其测量误差的方差通过检验具有齐次性, 就是平行测量。例如考试中的 A、B 卷。

(3) 多重记分的测验: 相对于二值记分的测验而言, 二值记分即答对记分, 答错不记分。而对一些由主观性题目(如语文考试中的作文, 英语考试中的写作)构成的测验记分可能是 0 到满分之间的任何一个分数, 这种记分方式的测验就叫多重记分的测验。

(4) 项目(item): 或称为题项, 即量表或试卷中的题目。

(5) 内在信度: 内在信度指的是量表中的一组问题(或整个量表)是否测量的是同一个概念, 即这些问题之间的内在一致性如何。如果内在信度系数在 0.8 以上, 则可以认为量表有较高的内在一致性。最常用的内在信度系数为克隆巴赫  $\alpha$  系数和折半信度。

(6) 外在信度: 指在不同时间进行测量时量表结果的一致性程度。最常用的外在信度指标是重测信度, 即用同一问卷在不同时间对同一对象进行重复测量, 计算一致程度。

## 3. 信度估计的方法

由于测验分数的误差来源不同, 估计信度的方法也有所不同。关于估计信度的具体方法请参见相关书籍, 在这里我们只针对 SPSS 中出现的信度估计方法进行介绍。请注意根据数据类型不同选择不同的方法。

### (1) $\alpha$ 信度系数

$\alpha$  信度系数是目前最常用的信度系数。它表明量表中每一题项得分间的一致性。该方法适用于项目多重记分的测验数据或问卷数据, 可以用该系数测量累加李克特量表(Likert-type Scale)的信度。累加李克特量表的数据输入格式见表 15-1。



其中， $x_{ij}$  ( $i=1, 2, \cdots, k; j=1, 2, \cdots, n$ ) 表示各受试对象第  $i$  个题目的得分，量表共有  $k$  个题目， $n$  名受试对象。 $\alpha$  信度系数公式为

$$\alpha = \frac{k}{k-1} \left( 1 - \frac{\sum_{i=1}^k S_i^2}{S_x^2} \right)$$

式中  $k$  为测验的题目数， $S_i$  为第  $i$  题得分数的方差， $S_x$  为测验总分的方差。

$\alpha$  信度系数可以解释用量表测试某一特质所得分数的变异中，有多大比例是由真分数所决定的，从而反映量表受随机误差影响的程度，即反映出测试的可靠程度。例如， $\alpha=0.90$  时，可以说测试所得分数的 90% 的变异是来自真分数的变异，仅有 10% 的变异来自随机误差。还可以把  $\alpha$  信度系数视作相关系数，它的取值范围从 0 到 1，出现负值是违反可靠性模型的。

用  $\alpha$  信度系数来估计量表的信度时，应注意  $\alpha$  信度系数与量表题目数量的多少有关。如一个含 10 个左右题目的量表，克隆巴赫  $\alpha$  系数应能达到 0.80 以上。如果题目增加，克隆巴赫  $\alpha$  系数会随之升高，条目多于 20 个时，克隆巴赫  $\alpha$  系数会很容易地升至 0.90 以上。如果量表的条目减少，克隆巴赫  $\alpha$  系数会随之降低。一个 4 个题目的量表，克隆巴赫  $\alpha$  系数有时可能会低于 0.60 或 0.50。因此，判断量表信度时，首先应当了解该量表题目的数量，然后再以此为基础，判断克隆巴赫  $\alpha$  系数是否达到了可以接受的水平。

(2) 分半信度

任何测验只是所有可能题目中的一份取样，如果抽取不同的部分，则可编制很多平行的等值测验，叫做复本（内容、形式相等的测验）。例如教师给学生出的 A、B 试卷。如果一种测验有两个以上的复本，根据一群被试接受两个复本测验的得分计算相关系数，即可得到复本信度，做可靠性分析。但建立复本是相当困难的，因此，在测验没有复本且只能实施一次的情况下，通常采用分半法估计信度，即测验题目分成对等的两半，根据各人在这两半测验的分数，计算其相关系数作为信度指标。其计算公式为

$$r_{xx} = \frac{2r_{hh}}{1+r_{hh}}$$

式中的  $r_{hh}$  为两半测验分数的相关系数， $r_{xx}$  为整个测验的信度估计值。应该注意的，如果测验的题目数量较少，比如 10 题以下，就不适合用这种方法来估计信度。

另外，分半法的使用基于人为分成两半的测验要等值，亦即两半测验的分数具有相同的平均数和标准差。当此条件不能满足时，就需要采用下面两个公式来估计信度。

① 弗朗那根公式

$$r = 2 \left( 1 - \frac{S_a^2 + S_b^2}{S_x^2} \right)$$

表 15-1 李克特量表数据结构

编号	问 卷 题 目				
	1	2	3	...	k
1	$x_{11}$	$x_{21}$	$x_{31}$	...	$x_{k1}$
2	$x_{12}$	$x_{22}$	$x_{32}$	...	$x_{k2}$
3	$x_{13}$	$x_{23}$	$x_{33}$	...	$x_{k3}$
...	...	...	...	...	...
n	$x_{1n}$	$x_{2n}$	$x_{3n}$	...	$x_{kn}$

式中  $S_a^2$  和  $S_b^2$  分别为两半测验分数的方差,  $S_x^2$  为测验总分的方差,  $r$  为信度值。

## ② 卢伦公式

$$r = 1 - \frac{S_d^2}{S_x^2}$$

式中  $S_d^2$  为两半测验分数之差的方差,  $S_x^2$  为测验总分数的方差,  $r$  为信度值。

## (3) 库德-理查逊 (Cuttman) 公式

倘若一个测验全由二值记分 (1, 0 方式记分) 的项目所组成,  $\alpha$  信度系数公式中每个项目上的分数方差就会等于该项目上通过率  $p$  与未通过率  $q$  两者的积。库德-理查逊公式为

$$r_{kk} = \frac{k}{k-1} \left( 1 - \frac{\sum p_i q_i}{S_x^2} \right)$$

其中  $k$  为构成测验的题目数,  $p_i$  为通过第  $i$  题的人数比例,  $q_i$  为未通过第  $i$  题的人数比例,  $S_x^2$  为测验总分的方差。

## (4) 平行测验的信度估计

对于信度, 也可定义为两平行测验上观察分数间的相关, 即用一个平行测验上某被试的观察分数, 去正确推论另一平行测验上该被试观察分数的能力, 用这种能力值的大小来定义测验的信度。平行测验信度估计的条件是方差具有齐次性, 有时还要求两平行测验的均数相等。

## 4. 数据要求与假设

(1) 数据要求: 用于分析的数据可以是数值型的二分数据、有序变量和间隔变量, 且为数值型。

(2) 假设: 观测量应该相互独立, 在各项目之间的误差应该互不相关。量表是可加的, 即各个项目得分相加即为总分数, 因此各个项目与总分数是线性相关的。

## 15.1.2 信度分析过程

1. 按 Analyze→Scale→Reliability Analyze 顺序打开如图 15-2 所示的主对话框。
2. 在左侧的源变量框中选择变量进入 Items 框, 作为分析变量。
3. 在源变量框下面有 Model 选项框, 用来选择估计信度系数的方法。单击向下箭头, 出现 5 种信度估计方法供选择。默认方法是 Alpha ( $\alpha$ ) 信度系数。

(1) Alpha,  $\alpha$  系数是内部一致性估计的方法, 适用于项目多重记分的测验 (主观题)。

(2) Split-half, 分半信度。将测验题分成对等的两半, 计算这两半分数的相关系数。

(3) Cuttman, 适用于测验全由二值 (1, 0) 方式记分的项目。

(4) Parallel, 是平行测验信度估计的方法, 条件是各个项目的方差具有齐次性。

(5) Strict Parallel, 除了要求各项目方差具有齐次性外, 还要求各个项目的均数相等。

4. 在 Scale Label 框内可以对所计算的信度系数进行说明, 例如在框内输入 “自觉

性维度的  $\alpha$  系数”。

5. 单击 Statistics 按钮, 打开相应对话框, 见图 15-3。在其中选择要输出的统计量。

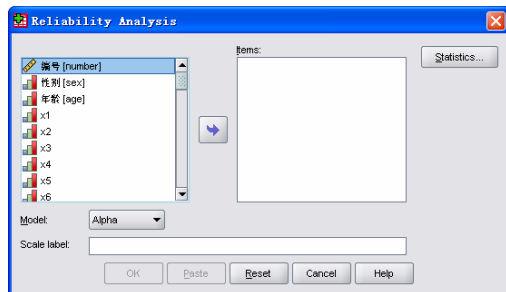


图 15-2 信度分析主对话框

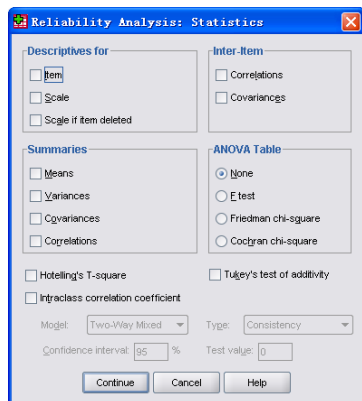


图 15-3 统计量选择对话框

### (1) Descriptives for 栏

① Item, 计算各项目的均数、标准差和样本含量。

② Scale, 计算量表的均数、标准差和项目数。即将各项目分数汇总得到总分数, 将总分数作为变量, 求其均数、标准差。

③ Scale if item deleted, 计算总分减去当前项目得分后计算的均数、方差等统计量。

### (2) Inter-Item 栏

① Correlations, 计算各项目间的相关系数。

② Covariances, 计算各项目间的协方差。

### (3) Summarizes 栏, 计算量表的描述统计量, 包括均值、方差、相关系数和协方差。

① Means, 对项目均数计算统计量, 包括项目均数的平均值、最小值、最大值、极差、最大值与最小值之比和项目均数的方差。

② Variances, 对项目方差计算统计量, 包括项目方差的平均值、最小值、最大值、极差、最大值与最小值之比和项目方差的方差。

③ Covariances, 对项目协方差计算统计量, 包括项目协方差的平均值、最小值、最大值、极差、最大值与最小值之比和项目协方差的方差。

④ Correlations, 对项目相关系数计算统计量, 包括项目相关系数的平均值、最小值、最大值、极差、最大值与最小值之比和项目相关系数的方差。

(4) ANOVA Table 栏中选择方差分析的方法, 是对均值相等的检验, 即检验各个项目上的得分是否具有—致性。

① None, 不生成方差分析表, 即不进行检验。这是系统默认选项。

② F Test, 输出重复测量方差分析表。

③ Friedman, 计算 Friedman 卡方值和 Kendall 谐和系数, 是对多个配对样本的平均

秩之间有无差异的检验。此时 Friedman 的卡方检验取代通用的  $F$  检验。

④ Cochran, 显示 Cochran's  $Q$  值。如果项目都是二分变量, 选择 Cochran。这时在 ANOVA 表中使用  $Q$  统计量取代常用的  $F$  统计量。它也是对多个配对样本的检验。

(5) Hotelling's  $T$ -square, 生成霍特林  $t^2$  统计量, 是对所有项目均数相等的零假设的多变量检验。

(6) Tukey's test of additivity, 给出量表提高可加性的功效估计值。检验假设是项目间没有交互作用。

(7) Intraclass correlation coefficient, 输出组内相关系数, 同时给出相关系数的置信区间、 $F$  统计量和显著性检验值。选中此项, 激活下面的选项。

① Model 选项框, 选择计算组内相关系数的模型。单击向下箭头, 有 3 种选择:

- Two-way mixed, 二维混合模型。
- Two-way random, 二维随机模型。
- One-way random, 一维随机模型。

② Type 选项框, 指定组内相关系数 (Intraclass Correlation) 是如何被定义的:

- Consistency, 选择此项, 表明研究中不关注评分者给出相同分数。
- Absolute Agreement, 选择此项, 表明研究者关注评分者给出相同的分数。

③ Confidence 框, 指定置信区间, 系统默认值 95%。

④ Test value 框, 为进行假设检验在此输入一个组内相关系数的假定值, 系统默认值是 0, 是相关系数为 0 的零假设。

6. 在主对话框选中 List item labels, 要求显示项目标号。

7. 在主对话框, 单击 OK 按钮, 提交运行。

### 15.1.3 信度分析实例

【例 1】本例题是心理学中研究运动员意志品质的调查问卷数据, 数据名 data15-01。问卷中有 50 个题目, 即 50 个项目。对 312 人进行了问卷调查。根据数据资料进行项目分析 (即对问卷做因子分析, 有关因子分析的内容, 请参见第 14 章) 后, 删除第 7、8、14、28、29、35、36、37、38、40、43、48 题, 并将剩余的 38 个项目根据项目分析的结果分为 5 个维度。5 个维度所包括的项目是:

自觉性维度:  $x_1$ 、 $x_2$ 、 $x_4$ 、 $x_{10}$ 、 $x_{13}$ 、 $x_{39}$ 、 $x_{41}$ 、 $x_{45}$ , 共 8 题。

果断性维度:  $x_{25}$ 、 $x_{31}$ 、 $x_{32}$ 、 $x_{34}$ 、 $x_{42}$ 、 $x_{44}$ 、 $x_{47}$ 、 $x_{49}$ 、 $x_{50}$ , 共 9 题。

自制力维度:  $x_3$ 、 $x_6$ 、 $x_{15}$ 、 $x_{17}$ 、 $x_{18}$ 、 $x_{21}$ , 共 6 题。

坚韧性维度:  $x_5$ 、 $x_9$ 、 $x_{11}$ 、 $x_{12}$ 、 $x_{16}$ 、 $x_{20}$ 、 $x_{23}$ 、 $x_{24}$ 、 $x_{26}$ 、 $x_{30}$ 、 $x_{46}$ , 共 11 题。

主动性维度:  $x_{19}$ 、 $x_{22}$ 、 $x_{27}$ 、 $x_{33}$ , 共 4 题。

表 15-2 为运动员意志品质评价量表 (预测版) 的格式。

现在要检验问卷的内部一致性如何, 即进行信度分析。用 SPSS 中的 Scale 信度分析

功能求得  $\alpha$  系数，就能说明该问卷中的 38 个变量的内部一致性结构如何。具体操作步骤如下：

表 15-2 运动员意志品质评价量表（预测版）

测 试 内 容	评 定 等 级				
	完全不符合	不太符合	说不清楚	比较符合	完全符合
1. 只要是一件有意义的事，我就会去做	1	2	3	4	5
2. 在一天的训练中，即使是重复同一动作，我也会一丝不苟地完成	1	2	3	4	5
3. 一件事完成以后，我经常后悔自己为何不早下决心	1	2	3	4	5
...	...	...	...	...	...
15. 对困难的任务我会想尽办法完成	1	2	3	4	5
16. 在以往比赛中，因为犹豫我错过了许多机会	1	2	3	4	5
...	...	...	...	...	...
24. 我决定做一件事时，常常是说干就干，决不拖拉或让它落空	1	2	3	4	5
25. 我常常为作决定犯难	1	2	3	4	5
...	...	...	...	...	...
27. 遇到棘手的事情我常常举棋不定，拿不出主意	1	2	3	4	5
...	...	...	...	...	...
32. 我主动找过教练商量下一步的练习计划	1	2	3	4	5
33. 如果见到有人落水，我会马上去救他	1	2	3	4	5

- 1. 建立数据集，见数据文件 data15-01。 $x_1 \sim x_{50}$  是问卷的题目，即项目。
  - 2. 按 Analyze→Scale→Reliability Analysis 顺序打开信度分析主对话框。
  - 3. 在主对话框的源变量框中，选中自觉性维度题项变量  $x_1$ 、 $x_2$ 、 $x_4$ 、 $x_{10}$ 、 $x_{13}$ 、 $x_{39}$ 、 $x_{41}$ 、 $x_{45}$ ，将它们选入 Items 框。
  - 4. 在源变量框下面的 Model 框内选择信度估计方法。本题用系统默认的 Alpha（ $\alpha$  系数）。
  - 5. 在 Scale Label 框内输入：自觉性维度的 Alpha 系数。
  - 6. 单击 OK 按钮运行，计算自觉性维度的信度系数  $\alpha$ 。
- 重复第 2 步，在主对话框单击 Reset 按钮，通过第 3、4、5、6 步骤分别对果断性、自制力、坚韧性、主动性维度计算信度系数。

表 15-3 量表各维度的信度系数表

项 目	信度分析表		
	样本量 (N of Cases)	项目数 (N of Items)	$\alpha$ (Alpha)
自觉性维度	312	8	0.144
果断性维度	312	9	0.353
自制力维度	312	6	0.271
坚韧性维度	312	11	0.460
主动性维度	312	4	0.042
量表	312	38	0.636

- 7. 主动性维度的信度系数求出后，在主对话框中将 38 个项目全部送入 Items 框，信度估计法还用系统默认的 Alpha 系数（ $\alpha$ ）。单击 OK 按钮运行。输出结果整理于表 15-3 中。
- 需要注意的是，如果问卷中有反向题（如果正向题给予 1、2、3、4、5 分，而反向题给予的是 5、4、3、2、1 分），则需要将其转换。
- 可以在输入数据时就直接将反向题进行转换。如果输入数据时未进行转换，可以单击 Transform→Recode into same variables 完成，详见第 2 章 2.3.3 节。本例题的数据已经对反向

题进行了转换。

### 8. 输出结果解释

表15-3中, 5个维度的信度系数分别为0.144、0.353、0.271、0.460、0.042, 而总量表的信度系数是0.636。

5个维度的信度系数都偏低, 需要进行问卷的修改。此外, 总量表的信度系数是0.636, 代表该量表的信度一般。如果要提高信度系数, 可以从5个维度中项目内容词句进行修饰、修改, 如果时间允许, 可增删项目, 再让这312名受试者测试一次。如果时间不许可, 在研究论文中应加以说明, 可作为今后研究的方向。

## 15.2 多维尺度分析(ALSCAL)

### 15.2.1 多维尺度分析的功能与数据要求

多维尺度分析(Multidimensional Scaling)是市场调查、分析数据的统计方法之一。通过多维尺度分析, 可以将消费者对商品相似性的判断产生一张能够看出这些商品间相关性的图形。例如, 有十个百货商场, 让消费者排列出对这些百货商场两两间相似的感知程度, 根据这些数据, 用多维尺度分析, 可以判断消费者认为哪些商场是相似的, 从而可以判断竞争对手。

1. 数据要求: 如果数据为不相似性数据, 它们必须为数值型数据或者是使用相同计量单位计量的数据。如果数据为多元变量, 数据可以是等间隔数据、二分数据或者是计数数据。注意应该保持数据量度单位的一致性, 否则将会影响到分析结果。如果不能避免这种情况的出现, 必须对数据进行标准化(在此分析过程中, 可以自动解决)。

2. 假设: 多维尺度分析没有严格的假设要求, 但在选择测量水平时应该十分小心。

### 15.2.2 多维尺度分析过程

1. 按 Analyze→Scale→Multidimensional Scaling (ALSCAL) 顺序打开多维尺度分析主对话框。如图 15-4 所示。

2. 在左侧的变量表中选择变量, 单击向右箭头, 将变量送入 Variables 变量框。

Variables 变量框下有 Individual Matrices for 框, 在此可输入一个变量进行分组。程序会为每一组分别计算距离矩阵, 同时无论是否在 Mode 对话框中选择了个体差异距离模型, 都会计算一个复本或加权距离模型。

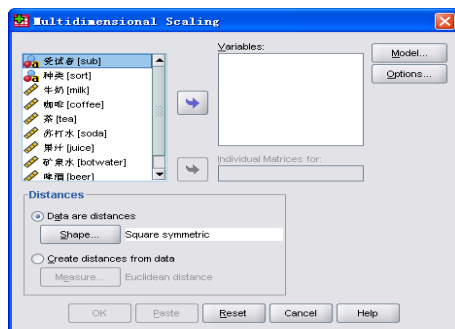


图 15-4 多维尺度分析主对话框

3. Distances 栏

(1) Data are distances, 当数据窗中的数据是一个或多个不相似性矩阵时选择此项。矩阵中的元素显示行和列两两配对的不相似程度。Shape按钮旁边显示的是当前选项。单击Shape按钮，打开数据形状框，见图15-5。

① Square symmetric, 方形对称结构。行、列代表相同的项目，且在上、下三角中相应的值相等。例如 A、B 两个项的相似性，A 与 B 和 B 与 A 的相似性是一样的，矩阵中上下三角对应位置上的值是相等的。

② Square asymmetric, 方形但不对称结构。行、列代表相同的项目，但上三角和下三角中相应的值是不相等的。如两个事物 A 与 B 比较和 B 与 A 比较所得分数不同。

③ Rectangular, 矩形结构。在 Number of rows 框中输入行数。在矩阵中行、列数据代表不同项目集。SPSS 把有序排列的数据文件当作矩形矩阵。如果数据中包含两个以上的矩形矩阵，一定要设定每个矩阵的行数。此数值必须大于等于 4。并且能够将矩阵中的行数整除（即各矩阵的行数应当相同）。

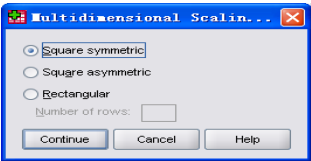


图 15-5 数据形状对话框

(2) Create distances from data, 根据数据生成距离阵。在 Measure 按钮之后显示的是当前选项。单击 Measure 按钮，打开数据测度方法对话框，见图 15-6。

① Measure 栏，可在此处选择用于分析的不相似性量度方法，方法说明参见附录 A。

② Create distance matrix 栏，创建距离矩阵。

- Between variables, 计算一对对变量之间的不相似性距离矩阵。
- Between cases, 计算两两观测量之间的不相似性距离矩阵。

③ Transform Value 栏，进行标准化转换，方法说明或算法参见附录 A。

4. 在主对话框单击 Model 按钮，进入 Model 对话框，如图 15-7 所示。在该对话框中确定数据和模型的类型。多维尺度分析的正确估计依赖于数据和模型。有以下选择栏：

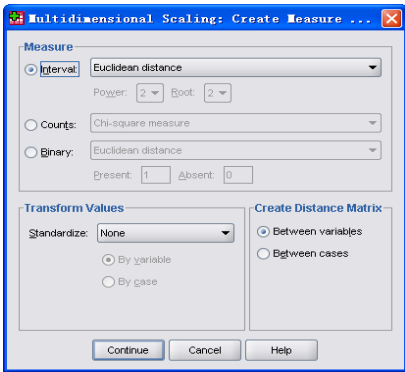


图 15-6 创建测量方法对话框

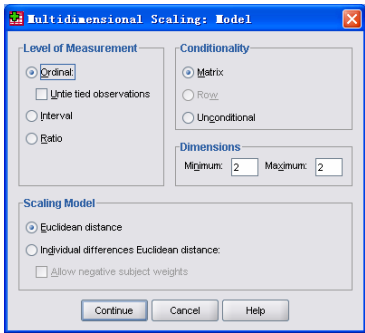


图 15-7 Model 对话框

(1) Level of Measurement 栏, 在该栏指定测度水平。有三个单选项:

- Ordinal, 有序测度的数据。若是有序分类数据, 使用 Kruskal 最小平方单调转换。指定 Untie tied observations, 对有相同分数的观测值赋予不同的秩。
- Interval, 数据是以间隔测度或定量的数据。
- Ratio, 以比例测度或定量的数据。

(2) Conditionality 栏, 在该栏指定模型类型, 有三个单选项:

- Matrix, 如果只有一个矩阵或每个矩阵代表不同的受试者时, 选择此项。
- Row, 只有行数据进行比较时有意义, 选择此项。且只适用于不对称或矩形矩阵。
- Unconditional, 矩阵内所有数值的比较都有意义时选择此项。

(3) Dimensions 栏, 用来指定多维尺度分析的维度。默认产生二维解。在 Minimum 框中输入最少维度数, 在 Maximum 中输入最多维度数。一般可选择计算 1~6 维度的解。为获得唯一解, 在最小和最大维度数中输入相同的数值。对于加权模型, Minimum 栏中至少应是 2。

(4) Scaling Model 栏, 指定尺度模型, 有两个单选项:

• Euclidean distance, 欧几里德模型可以应用于任何类型的矩阵分析中。如果数据中只包含一个矩阵, 那么将进行 CMDS 典型多维尺度分析; 如果包含两个以上的矩阵, 进行 RMDS 重复多维尺度分析。

• Individual differences Euclidean distance, 加权个体差异欧几里德距离模型 (WMDS)。该模型需要两个或以上的矩阵。

5. 单击 Options 按钮, 进入 Options 对话框, 见图 15-8。

(1) Display 栏, 在该栏中选择输出项。

① Group plots, 多维尺度分析图。这个图在多维尺度分析中非常重要。可以利用这个图对每一维寻找散点间相关性的合理的解释。

② Individual subject plots, 对有序分类数据或模型中指定 Matrix 的数据显示每一个受试者的图形, 而对模型中指定 Row 的数据无效。

③ Data matrix, 显示每一个受试者的数据矩阵。

④ Model and options summary, 输出所有选项的基本信息, 包括数据选项、模型选项、输出选项和迭代判据选项等信息。

(2) Criteria 栏设置迭代停止的判据, 有三个选项:

① S-stress convergence 参数框, 单调收敛准则, 系统默认在拟合距离模型过程中计算拟合劣度指标 S-stress。当从一个迭代到下一个迭代的 S-stress 变化量 (即拟合的改善量) 等于或小于 0.001 时, 迭代停止。为了提高解的精度, 可以输入一个比以前设置值小的正

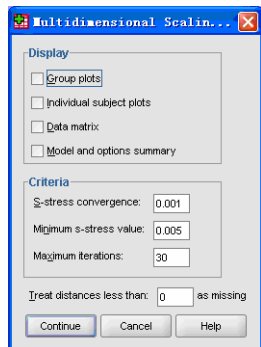


图 15-8 Options 对话框



值。如果输入零，只进行30步的迭代。

② Minimum s-stress value参数框，最小s-stress值，系统默认收敛值为0.005时迭代停止。如果要继续进行迭代，输入一个比默认值更小的数值。如果输入的数值比默认值大，迭代次数会减少。该值要大于0，小于或等于1。

③ Maximum iterations参数框，用最大迭代次数作为迭代停止的判据。当最大迭代次数等于设置值时迭代停止。系统默认值是30。如果输入值比默认值大，增加分析的精度，但计算时间也会增加。

(3) Treat distances less than *n* as missing,系统默认将距离小于 0 的值作为缺失值。用户可以指定 *n* 值，系统把小于 *n* 的值为缺失值处理。

6. 在主对话框，单击 OK 按钮，执行操作，系统将输出多维尺度分析结果。

15.2.3 多维尺度分析实例

【例 2】本例题使用《SAS 系统与市场调查数据分析》（高慧璇等编著）一书中的例题数据。该数据是假设 7 名受试者按照 1~7 的尺度（1 表示非常相似，7 表示非常不相似）排列出一些饮料间两两相似的感知程度。这些饮料作为变量包括：milk（牛奶）、coffee（咖啡）、tea（茶）、soda（苏打水）、juice（果汁）、botwater（矿泉水）、beer（啤酒）、wine（葡萄酒）。要求受试者给出这些饮料的两两相似的感知程度，共有 28 种可能  $(n(n-1)/2)$ 。用此数据分析哪些饮料消费者认为是相似的。该分析可以使用多维尺度分析（ALSCAL）方法完成。具体步骤如下：

sub	sort	milk	coffee	tea	soda	juice	botwater	beer	wine
sub1	milk	1	6	7	7	7	7	7	7
sub1	coffee	6	1	1	7	7	7	7	6
sub1	tea	6	1	1	7	5	4	7	5
sub1	soda	7	7	7	1	5	3	5	4
sub1	juice	7	7	5	5	1	5	3	2
sub1	botwater	7	7	4	3	5	1	6	6
sub1	beer	7	7	7	5	3	6	1	1
sub1	wine	7	6	5	4	2	6	1	1
sub2	milk	1	5	7	7	7	7	3	7
sub2	coffee	5	1	6	7	7	6	7	7
sub2	tea	7	6	1	6	4	4	3	7
sub2	soda	7	7	6	1	7	7	7	4
sub2	juice	7	7	4	7	1	4	6	4
sub2	botwater	7	6	4	7	4	1	7	7
sub2	beer	3	7	3	7	6	7	1	5
sub2	wine	7	7	7	4	4	7	5	1

图 15-9 data15-03 数据文件结构

1. 打开数据文件，结构见 data15-02。因为本例题的数据矩阵是对称的，例如牛奶与咖啡间的距离和咖啡与牛奶间的距离一样，所以可以做成三角矩阵。sub 变量为受试者编号。每个受试者对 7 种饮料两两比较，根据它们之间的相似度打分，7 分制。比较中如果所给分值越大表明相似程度越高，则定义为相似数据。比较中如果所给分值越小表明相似程度越高，则定义为不相似数据。例如如图 15-9 所示的数据就是不相似数据，第一个受试者认为牛奶与牛奶非常相似，两者的相似度打分为 1；咖啡与牛奶不相似，认为两者的相似度为 6，以此类推。每个受试者的数据是一个矩阵，其数据结构见图 15-9。

2. 按 Analyze→Scale→Multidimensional Scaling（ALSCAL）顺序打开多维尺度分析的主对话框。

3. 选中分析变量 milk、coffee、tea、soda、juice、botwater、beer、wine，送入分析

变量框。注意输入分析变量的顺序一定要与数据文件中的顺序一致。

4. 在 Distances 单选框内选中 Data are distance 项, 单击 Shape 按钮, 展开对话框。

5. 在 Shape 对话框中选中 Square symmetric 项, 因为本例数据的行与列项目相同, 上三角与下三角的值是相同。

6. 在主对话框, 单击 Model 按钮, 展开 Model 对话框。

(1) Level Measurement 栏, 因为用 1~7 给饮料的相似度评分, 所以选 Ordinal。

(2) Scaling Model 栏选 Euclidean distance 项要求拟合欧几里得距离模型。

(3) Conditionality 栏选 Matrix 项。因为每个矩阵代表一个被试的答案。计算二维解, 故在 Dimensions 栏的 Minimum 和 Maximum 项均输入 2。

7. 在主对话框单击 Options 按钮, 进入 Options 对话框。

(1) 在 Displays 栏选 Group Plots 项, 做多维尺度分析图。

(2) 在 Criteria 栏使用系统默认判据。

8. 输出结果, 见图 15-10~图 15-13 和表 15-4。最关注的是多维尺度分析图(有的参考书称之为共用感知图)。

### 9. 结果解释

图 15-10 给出了二维结果的迭代的过程。在 Criteria 栏指定的迭代最大数为 30, 但当拟合劣度 S-Stress 的改善值小于 0.001 时迭代终止。本例迭代到第四步时 S-Stress 的改善值是 0.00062, 其值小于 0.001, 迭代过程结束。

图 15-11 给出了 Stress 和 RSQ 值。RSQ 即  $R^2$ , 它是拟合优度指标, 数值越接近 1, 表明模型拟合越好; Stress 是拟合劣度指标, 百分比值越大说明模型拟合越差。表 15-4 给出了 Stress 大小与拟合好坏的一个参考。

本例题的 Stress 值为 0.30437 (30.4%), RSQ 值是 0.37281, 表明模型拟合的不好。解决方法, 一个是用近似多维尺度分析 PROXSCAL 方法, 另一个是再增加受试者。

Iteration history for the 2 dimensional solution (in squared distances)		
Young's S-stress formula 1 is used.		
Iteration	S-stress	Improvement
1	.45654	
2	.41326	.04327
3	.40999	.00326
4	.40936	.00062
Iterations stopped because S-stress improvement is less than .001000		

Stress and squared correlation (RSQ) in distances					
RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula 1.					
Matrix	Stress	RSQ	Matrix	Stress	RSQ
1	.285	.450	2	.375	.045
3	.318	.322	4	.247	.562
5	.147	.851	6	.343	.195
7	.354	.164			
Averaged (rms) over matrices					
Stress = .30437		RSQ = .37281			

表 15-4 拟合量度值评价

Stress (%)	拟合度
20	差
10	一般
5	好
2.5	较好

图 15-10 二维解决方案迭代过程

图 15-11 Stress 和 RSQ 值

图 15-12 给出了二维导出构形表, 表中的数值是用在多维尺度分析图的坐标值。

图 15-13 为多维尺度分析图。该图是我们进行多维尺度分析最关注的结果。从此图中可解释的内容包括: 对图形的每一维寻找散点间相关性的合理解释。从图中可看出, 包括三组聚焦点, 这意味着消费者认为彼此相似的这些产品: 咖啡和茶是相似的, 果汁

和牛奶是相似的，啤酒和葡萄酒是相似的。说明这些相似饮料在市场占有率上彼此有竞争。另外，从垂直维 Dimension 2 看，可将七种饮料分为两类，牛奶、果汁、苏打水和矿泉水属于营养型饮料，啤酒、葡萄酒、咖啡和茶属于提神型饮料。

Derived Stimulus Configuration  
Euclidean distance model

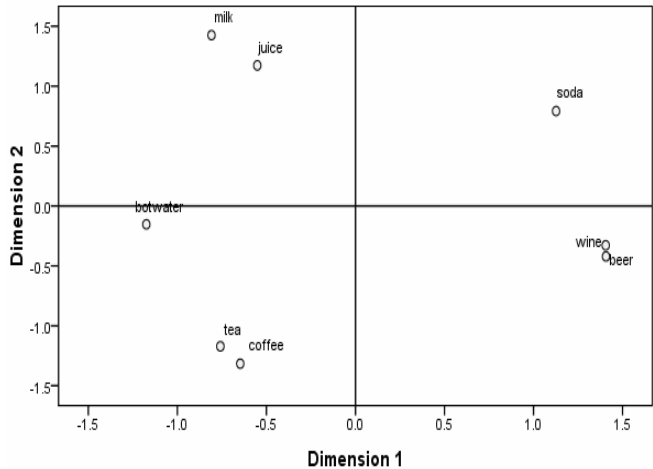


图 15-12 二维导出构形表

图 15-13 多维尺度分析图

习 题 15

1. 信度分析中用哪些指标可以反映问卷可靠性？如果要了解问卷中的某一个维度的可靠性程度如何，应怎么做？
2. 反映量表内部一致性高低的克隆巴赫（ $\alpha$ ）系数与量表题目的数量有关吗？
3. 什么是相似数据，什么是不相似数据，多维尺度分析的目的是什么？
4. data15-03 是一个受试者对牙膏认识的数据。试进行不相似数据的多维尺度分析。

# 第 16 章 结 合 分 析

## 16.1 结合分析概述

### 1. 结合分析的概念

结合分析是测度消费者对产品属性各侧面或售后服务等的偏好的一种技术。

在市场调查中经常想了解顾客对产品的偏好，作为产品销售策略制定的依据或者作为新产品研制决策的依据。

每个作为商品的产品都是由一系列的属性和服务构成的。例如一台计算机的属性有 CPU、显示器、内存、硬盘、品牌、价格以及售后服务等。每个属性描述了一台计算机的一个侧面。在计算机技术发展的一个特定时期，每个属性都有几个技术指标。例如 CPU 有单核、双核、三核、四核之分，显示器有 14 英寸、15 英寸、17 英寸、21 英寸等，硬盘有 100GB、500GB、800GB 及 1TB 的等，内存有 256MB、512MB、1GB、2GB 等。每个属性就是顾客购买决策所考虑的因素，每个属性的每个技术指标就是这个因素的一个水平。

市场研究中的顾客偏好调查分析要求被访者对各种属性水平的组合给出自己的偏好得分，或者将各种组合排序，给出各种组合的秩。这些得分或者排序后的秩分就是对顾客偏好的测度。结合分析依据这些数据分析判断得出顾客偏好的结论。

### 2. 结合分析的步骤

这里阐述的所要求的步骤与 SPSS 的结合分析程序有关。

(1) 分析产品的属性，确定每个属性的水平数和水平的具体内容。在统计分析中也称属性为因素。

应该选择课题研究的主要因素。选择有代表性的重要的属性是偏好分析的重要环节。属性的数量应该尽量精简。每个属性的水平数也应该在达到课题要求的前提下，尽量少。

### (2) 实验设计

将选择的属性（因素）水平组合成实验组，每个组的因素水平组合称为产品的一个侧面。它是要呈现在被访者面前，供被访者评价的。为减小误差，节省人力、物力、时间，使调查更加有效，通常采用正交设计。可以使用 SPSS 的正交设计程序产生要求数目的侧面，也可以由读者输入形成设计文件。

SPSS 的正交设计程序产生设计文件供调查使用，同时也是结合分析的必要数据。

### (3) 根据设计打印调查卡片

### (4) 运用各种调查方法取得数据

调查取得的数据有两种:

- ① 要求被访者对所设计的侧面排秩, 如最喜欢的秩为 1, 次之的为 2, 以此类推;
- ② 要求被访者为所设计的侧面打分, 如最喜欢的分数为 100, 最不喜欢的分数最低。

### (5) 程序设计与运行

SPSS 中没有窗口式的结合分析程序, 必须使用 SPSS 语句进行程序设计, 运行设计的程序分析调查得到的数据, 在输出窗得到输出结果。

(6) 根据输出结果选择顾客最偏爱的产品属性组合, 作为开发新产品的决策依据, 或制定销售策略的依据。

### 3. 本章用到的术语

(1) 侧面, 是指所研究的产品属性(因素)水平的组合, 在正交设计中产生。

(2) 全概念侧面, 是指能代表各种属性的全部组合, 正交设计结果中的侧面可称为全概念侧面, 用此进行偏好调查, 分析结果将是可信的决策依据。

(3) 实验侧面, 即要打印成卡片或出现在调查问卷中、由被访者评价的侧面。它是由正交设计形成的。

(4) 保留侧面, 是正交设计侧面以外的, 为进行对估计效应有效性的检验而建立的侧面。保留侧面由另一个随机设计产生, 不是由正交设计产生的。

(5) 模拟侧面, 由读者输入的侧面。

(6) 设计文件, 由正交设计过程生成或者由读者输入, 符合正交性的数据文件。文件中的变量就是课题确定的感兴趣的因素, 是所研究产品的一个属性。而观测量是由各变量水平值组成的产品的侧面。它是一个各因素水平的组合。

设计文件还可以包括保留侧面和模拟侧面。这两个侧面应该由一个特殊变量 `Status_` 标识。保留侧面和模拟侧面的 `STATUS_` 值分别为 1、2。实验侧面的 `STATUS_` = 0。

## 16.2 正交实验设计

### 16.2.1 实验设计中的问题

众所周知, 在调查中, 产品的属性数和各属性的水平数不能太多, 否则其组合数就会很大, 以至于调查和获取数据简直就是不可能完成的任务。例如 2 个属性, 每个属性取 3 个水平, 就有 9 种组合。如果每个属性有 5 个水平, 就会有 25 种组合; 如果有 5 因素, 每个因素 3 水平, 组合数是 243 ( $3 \times 3 \times 3 \times 3 \times 3$ )。要求顾客对 243 种产品打分、排序肯定得不到很好的结果。因此首先要选择有代表性的属性和水平, 而且要有效地减少调查中呈现在调查对象面前的组合。

因此进行实验设计时要考虑:

(1) 当要调查的产品属性(因素)不只一个, 而且, 每个属性的水平也不只一个时, 要合理安排各个属性水平组合, 以便降低由于被访者对组合理解的差异所引起的误差。

- (2) 以最少的属性（因素）水平组合数进行调查，得到可靠的结论。
- (3) 节省人力、物力、财力和时间。
- (4) 便于使用软件进行结合分析，提高调查对顾客偏爱估计的准确性。

### 16.2.2 正交实验设计的思路

为简化问题，现在以三因素 2 水平的实验设计为例。

为调查酸奶饮品的顾客偏爱。选择酸奶的品牌、直接原料、附加成分为产品调查的侧面，每个侧面选择两个水平。要调查哪种水平组合是顾客最爱的。

三因素及其水平表示：因素用大写字母表示，A（品牌）、B（直接原料）、C（成分）。可以看做一个三维坐标系。见图 16-1。

各因素的水平序列号用跟在因素字母后的阿拉伯数字表示：

A 因素，品牌的 2 个水平为：A1（三元）、A2（伊利）

B 因素，直接原料的 2 个水平为：B1（鲜牛奶）、B2（纯牛奶）

C 因素，附加成分的 2 个水平为：C1（VAD）、C2（高钙）

全面实验，即各因素的各水平全部组合一次，有  $2^3$  共 8 次实验，见表 16-1。这些组合可以用正方体的 8 个顶点表示，见图 16-2。这些组合是 A1B1C1, A2B1C1, A1B1C2, A2B1C2, A1B2C1, A2B2C1, A1B2C2, A2B2C2。

表 16-1 全面实验组合表

		A1	A2
B1	C1	A1B1C1	A2B1C1
	C2	A1B1C2	A2B1C2
B2	C1	A1B2C1	A2B2C1
	C2	A1B2C2	A2B2C2

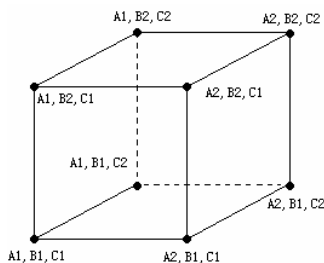
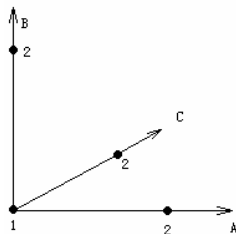


图 16-1 三因素各有两个水平      图 16-2 全面实验组合示意

再看一看图 16-3，只取 4 种水平组合就可以代替上述的 8 种全面实验组合。这样就简化成 4 个实验：A1B1C1、A2B1C2、A1B2C2、A2B2C1，见表 16-2。

为什么 4 次实验可以代替 8 次实验呢？

1. 观察正方体，三个因素的每个水平都均匀地包含在这四个组合中了。选中的四个组合均匀地分布在正方体中。每个面都有两个点，每个线都有一个点。分布均匀。
2. 再观察正交表 16-3，具有下列特点：

(1) 每个因素（列）的每个水平都出现，且出现的次数相同：1 水平出现两次，2 水平出现两次。

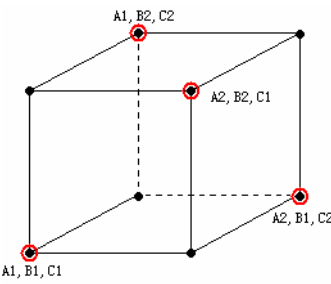


图 16-3 正交设计示意图

表 16-2 4 个顶点的组合

	A1	A2
B1	C1	C2
B2	C2	C1

表 16-3 正交设计结果表

列号 实验号	A	B	C
1	1	1	1
2	2	1	2
3	1	2	2
4	2	2	1

(2) 任意两个因素（任意两列）的水平数据对是相同的。正交表中 A、B 两列的水平搭配是(1,1)，(2,1)，(1,2)，(2,2)；A、C 两列的水平数据对是(1,1)，(2,2)，(1,2)，(2,1)与 A、B 两列是相同的只是顺序不同。所以因素水平的搭配是均匀的。

具有上述特点的实验设计表就称为正交表。

以上两个性质称为正交表的正交性。这个性质使得正交表在使用部分组合进行实验时具有以下特点：

(1) 正交表中列出的实验组合能很好地代表全面的实验组合：

第一个性质各因素各水平在每列中都出现相同的次数，保证了这些实验组合对全面实验的代表性。

第二个性质使任意两个因素间的组合为全面的实验组合从而保证了而且使部分实验找到的最优组合与全面实验的结果趋势一致。

(2) 实验组合均衡地分布在全面实验组合之中，见图 16-3。

(3) 正交性使得任一因素各水平的实验条件相同。这就保证了在每列因素各水平的效果中，最大限度地排除了其他因素的干扰。从而可以综合比较该因素不同水平对实验指标的影响情况。

根据上述正交设计的结果，在调查问卷中呈现在被访者面前的是这样的牛奶的属性组合：

- ① 三元牌 VAD 鲜牛奶
- ② 伊利牌 高钙 鲜牛奶
- ③ 三元牌 高钙 纯牛奶
- ④ 伊利牌 VAD 纯牛奶

16.2.3 正交实验设计过程

从主菜单中逐一选择：Data→Orthogonal Design→Generate 打开正交设计主对话框，如图 16-4 和图 16-5 所示。

1. 定义因素和因素水平，步骤如下：

(1) 在 Factor Name 栏中输入因素变量名，必须是合法的 SPSS 变量名，但不能用 Status\_ 和 Card\_ 作为因素变量名。

(2) 在 Factor Label 栏中输入因素变量的标签。

(3) 单击 Add 按钮将因素变量名及其标签，转入到大矩形框中。显示格式为：因素变量名“标签”[?]。

可以重复上述三步操作，定义若干个因素变量。

如果要修改因素变量名和标签的定义，先选择它，因素变量名和标签重新返回 Factor Name 栏和 Factor Label 栏。在这两个栏中修改后，单击被激活的 Change 按钮。修改后的因素变量名和变量标签显示在大矩形框中。如果要删除已经定义的因素变量及其标签，只要在选择它后，单击 Remove 按钮即可。

(4) 从大矩形框中选择一个因素变量，激活 Define Values 按钮。展开相应的二级对话框，如图 16-6 所示。

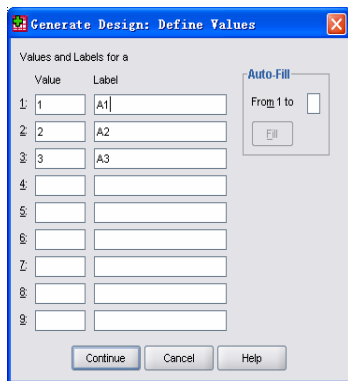
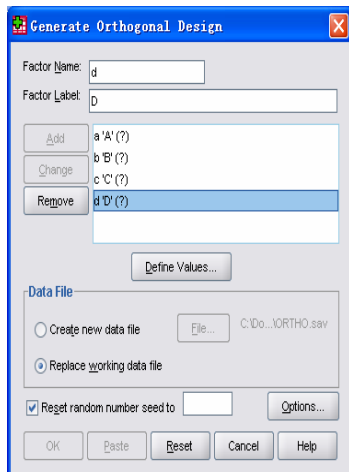
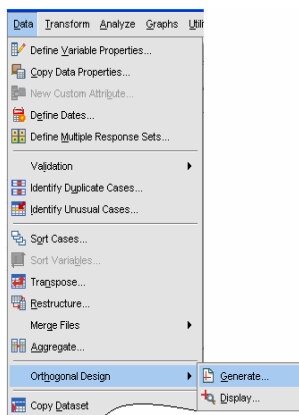


图 16-4 菜单选择

图 16-5 正交设计主对话框

图 16-6 定义因素水平对话框

(5) 在二级对话框中定义因素变量的值和值标签。

① 如果水平较多，且水平值是从 1 开始的，可以利用 Auto-Fill 栏中的自动功能，自动填入因素水平值。方法是将水平数输入到 From 1 to 后面的文字编辑框中，单击 Fill 按钮，例如图 16-6 中填入 3，单击 Fill 按钮后，在 Value 栏中自动填入 1、2、3 三个水平值。

② 将各水平值的含义，作为标签输入到对应水平值后边的 Label 列的单元格中。

③ 单击 Continue 按钮返回主对话框。重复 (4)、(5) 两步骤，将所有因素变量的值标签定义工作完成。

2. 定义设计结果保存方式。在 Data File 栏内根据保存要求，选择保存方式：



(1) Create new data file, 把设计结果保存到一个数据文件中。选择此项, 激活 File 按钮, 并在其后显示默认的保存位置和默认的数据文件名。默认的保存位置为当前目录, 默认的文件名为 Ortho.sav, 扩展名 sav 表明默认的文件类型是 SPSS 数据文件。

如果想改变保存位置和文件名, 单击 File 按钮, 打开 Output File Specification 对话框。指定你自己的保存位置、文件类型和文件名, 单击“保存”按钮, 返回主对话框。

(2) Replace working data file 代替当前的数据文件。

3. 选中 Reset random number seed to 复选项后, 在其后的框中重新为随机数种子指定一个值。在生成正交设计过程中, 要通过随机数种子, 产生随机数。相同的随机数种子产生相同的设计结果, 因此不同的设计, 要设置不同的种子值。必须在生成第一个设计之前设置该种子值, 种子值可以是 1 到 2 000 000 000 中的任意整数。在一个 SPSS 执行周期中, 如果想生成几个相同的随机数集, 并在后续的设计生成时, 将种子值再次设置成相同的值。

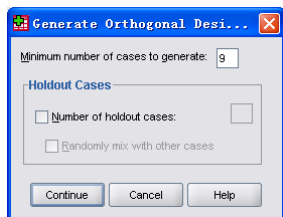


图 16-7 正交设计的选项

4. 正交设计生成程序的选项。单击 Options 按钮打开相应的对话框, 见图 16-7。

(1) Minimum number of cases to generate 框, 指定一个计划设计要生成的最少观测量数。即指定一个实验设计的最小实验数。选择一个正整数, 要小于等于根据所有因素水平可能的组合构成的观测量总数, 也就是要小于等于全模型的实验数。

如果不明确指定要生成的观测量最小数, 则自动生成对正交设计必要的数量的观测量。如果 Orthoplan 过程不能产生至少是大致所要求的最少观测量数, 就将产生符合所指定的因素和水平数的最大数。注意: 该设计没有必要包括确切的指定的观测量数。但使用这个值作为最小值在正交设计中生成更合适的, 可能是最小观测量数。例如, A 因素 3 水平、B 因素 2 水平、C 因素 2 水平。总组合数是 12。设置要生成的最小观测量数必须小于或等于 12。

(2) Holdout Cases 栏设置有关产生除正规设计的观测量以外的保留观测量的选项。

① Number of holdout cases, 设置除正规设计外的观测量数。但 Conjoint 过程估计效应时不使用这些额外的观测量。在被激活的文本编辑框中输入一个正整数, 该数值必须小于等于由因素水平组合决定的观测量总数。

② Randomly mix with other cases, 输出结果随机地将保留观测量与实验观测量混合。如果不选择这个选项, 保留观测量在数据文件中出现在实验观测量后面。

③ 单击 Continue 按钮, 返回主对话框。

5. 主对话框中单击 OK 按钮, 提交系统执行, 输出的设计结果保存到指定位置, 并在输出窗口给出可能条件组合的设计结果。

### 16.2.4 正交实验设计实例

【例1】要求生成4因素3水平9次实验的正交实验设计表。

1. 操作步骤:

(1) 按 Data→Orthogonal Design→Generate 顺序单击菜单项, 展开正交设计对话框。

(2) 在主对话框中定义4个因素变量, 变量名为  $a$ 、 $b$ 、 $c$ 、 $d$ , 变量标签分别为变量名相应的大写字母 A、B、C、D。

(3) 逐个选择因素变量, 单击 Define Values 按钮在相应的对话框中定义因素水平值及其值标签:

$a$  [A]: A1、A2、A3      $b$  [B]: B1、B2、B3

$c$  [C]: C1、C2、C3      $d$  [D]: D1、D2、D3

(4) 在 Data File 栏内选择 Replace Working data file, 设置将设计结果显示在工作数据文件中, 即当前的数据窗口中。

(5) 设置随机数种子, 随便填写一个正整数: 2345。

(6) 单击 Options 按钮, 打开相应对话框。在 Minimum number of cases to generate 后面的矩形框中输入数字9。单击 Continue 按钮返回主对话框。

单击 Paste 按钮生成程序如下:

```
SET SEED 2345.
```

```
ORTHOPLAN
```

```
/FACTORS = a 'A' ( 1 'A1' 2 'A2' 3 'A3') b 'B' ( 1 'B1' 2 'B2' 3 'B3') c 'C' ( 1 'C1' 2 'C2' 3 'C3') d 'D' ( 1 'D1' 2 'D2' 3 'D3')
```

```
/REPLACE /MINIMUM 9.
```

2. 在工作数据窗口中生成正交设计结果。见图 16-8 (a)。改变随机数种子为 5678, 结果见图 16-8 (b)。两个设计结果是不同的。

如果在 DATA File 栏选择了 Create New File, 并指定了保存位置, 生成的设计保存在指定位置的数据文件中。

```
*Generate Orthogonal Design.
```

```
SET SEED 2000.
```

```
ORTHOPLAN
```

```
/FACTORS=A 'AA' (1 'a1' 2 'a2' 3 'a3') B 'BB' (1 'b1' 2 'b2' 3 'b3') C 'CC' (1 'c1' 2 'c2' 3 'c3') D 'DD' (1 'd1' 2 'd2' 3 'd3')
```

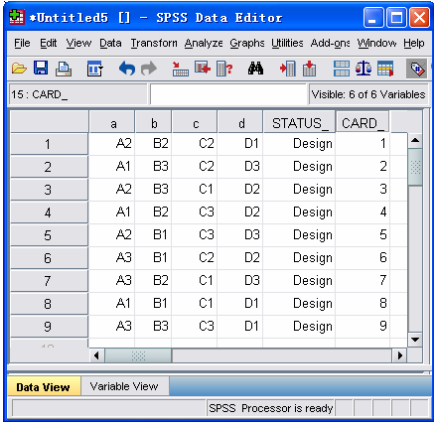
```
/OUTFILE='E:\书\SPSS 统计分析第4版\data\正交设计例 1.sav'
```

```
/MINIMUM 9 /HOLDOUT 3 /MIXHOLD YES.
```

可以按照当前工作数据窗口中的实验设计结果, 安排组织实验了。

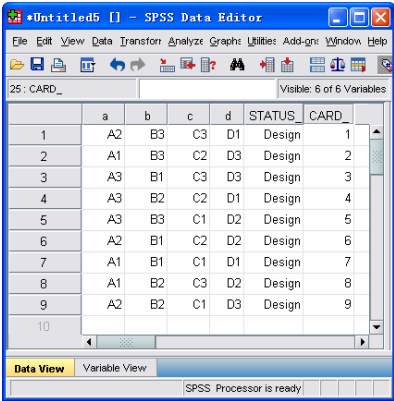
如果想要多做两次实验, 在 Options 对话框的 Holdout Cases 栏中选择 Number of

holdout cases，并输入 2。最后的运行结果见图 16-9（a）。图 16-9(a)比图 16-8(a)在最后多出两个观测量，即第 10、11 行。其 Status\_变量的值为 Holdout。



	a	b	c	d	STATUS_	CARD_
1	A2	B2	C2	D1	Design	1
2	A1	B3	C2	D3	Design	2
3	A2	B3	C1	D2	Design	3
4	A1	B2	C3	D2	Design	4
5	A2	B1	C3	D2	Design	5
6	A3	B1	C2	D3	Design	6
7	A3	B2	C1	D3	Design	7
8	A1	B1	C1	D1	Design	8
9	A3	B3	C3	D1	Design	9

(a)

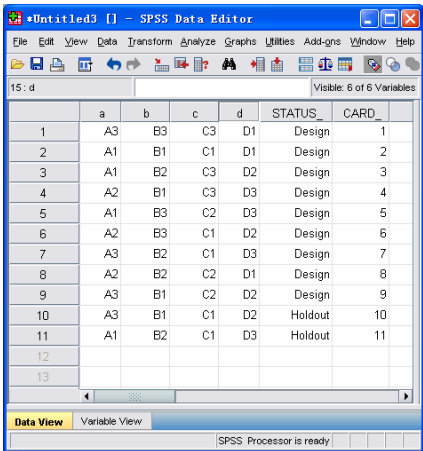


	a	b	c	d	STATUS_	CARD_
1	A2	B3	C3	D1	Design	1
2	A1	B3	C2	D3	Design	2
3	A3	B1	C3	D3	Design	3
4	A3	B2	C2	D1	Design	4
5	A3	B3	C1	D2	Design	5
6	A2	B1	C2	D2	Design	6
7	A1	B1	C1	D1	Design	7
8	A1	B2	C3	D2	Design	8
9	A2	B2	C1	D3	Design	9

(b)

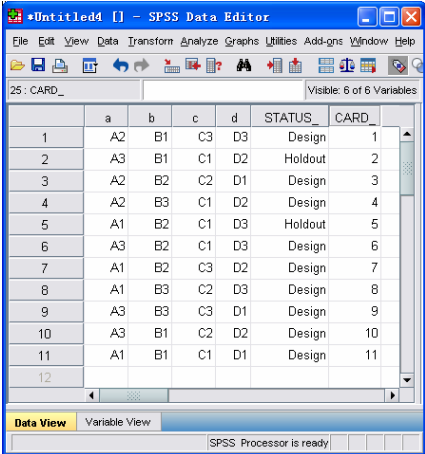
图 16-8 四因素三水平九次实验的正交实验设计结果

如果选择了 Number of holdout cases，并输入 2 的同时，还选择了 Randomly mix with other cases，输出结果随机地将维持观测量与实验观测量混合。见图 16-9(b)。Status\_变量值为 Holdout 的观测量随机地混在原正交设计的观测量中。



	a	b	c	d	STATUS_	CARD_
1	A3	B3	C3	D1	Design	1
2	A1	B1	C1	D1	Design	2
3	A1	B2	C3	D2	Design	3
4	A2	B1	C3	D3	Design	4
5	A1	B3	C2	D3	Design	5
6	A2	B3	C1	D2	Design	6
7	A3	B2	C1	D3	Design	7
8	A2	B2	C2	D1	Design	8
9	A3	B1	C2	D2	Design	9
10	A3	B1	C1	D2	Holdout	10
11	A1	B2	C1	D3	Holdout	11
12						
13						

(a)



	a	b	c	d	STATUS_	CARD_
1	A2	B1	C3	D3	Design	1
2	A3	B1	C1	D2	Holdout	2
3	A2	B2	C2	D1	Design	3
4	A2	B3	C1	D2	Design	4
5	A1	B2	C1	D3	Holdout	5
6	A3	B2	C1	D3	Design	6
7	A1	B2	C3	D2	Design	7
8	A1	B3	C2	D3	Design	8
9	A3	B3	C3	D1	Design	9
10	A3	B1	C2	D2	Design	10
11	A1	B1	C1	D1	Design	11
12						
13						

(b)

图 16-9 四因素三水平九次实验带有 2 个保留观测量的正交实验设计结果

16.2.5 正交设计过程语句

1. 正交设计过程 ORTHOPLAN 使用下列语句调用

ORTHOPLAN [FACTORS=varlist ['labels'] (values ['labels'])...]

```
[{/REPLACE }] [/OUTFILE='savfile'|'dataset'] [/MINIMUM=value]  
[/HOLDOUT=value] [/MIXHOLD={YES}]{NO }
```

其中, ORTHOPLAN是命令关键字。ORTHOPLAN命令为结合分析产生正交的主效应设计。设计结果可以添加在当前的工作数据集中, 如果没有工作数据集存在也可以建立一个工作数据集。产生的全组合设计可以列出或使用PLANCARDS格式安排。由ORTHOPLAN产生的文件可以用作CONJOINT命令要求的设计文件。

## 2. 基本要求

ORTHOPLAN命令关键字后面跟着FACTORS及等号后面的变量列表, 如果有变量标签, 放在每个变量名后面的引号中。

每个变量名、变量标签后面列出该变量的水平值, 如果有值标签, 放在值后面的引号中。每个变量的值和值标签列表放在变量名和变量标签后面的括号中。

ORTHOPLAN在工作数据集中产生观测量。用每个观测量描述结合实验设计的一个侧面, 由因素值组合而成。默认是生成最小可能的正交设计。

如果已经存在工作数据集包括正交设计的所有变量, FACTORS就是可选的子命令。

## 3. 子命令功能与限制

子命令在ORTHOPLAN命令语句后面, 出现的顺序任意。

### (1) 有关的运行结果

① 如果原来没有工作数据集, ORTHOPLAN 通过 FACTORS 子命令, 使用变量和变量值信息建立工作数据集。

② ORTHOPLAN 产生的数据附加在一个工作数据集上。如果没有使用 FACTORS 子命令, 因素水平值必须在前面一个 ORTHOPLAN 或 VALUE LABELS 命令定义。

③ 新变量 STATUS\_ 和 CARD\_ 如果原来不存在, 就产生并附加在 ORTHOPLAN 产生的工作数据集中。实验观测量的 STATUS\_ 变量值为 0, 保留观测量的 STATUS\_=1, 模拟观测量的 STATUS\_=2。保留观测量由被访者做评价, 但是在结合分析 CONJOINT 的效应估计中不用。而是用这些观测量作效应估计的合法性检验。模拟观测量由用户输入。它们是不由被访者评价的因素水平组合, 但是以实验观测量的评价为基础由 CONJOINT 进行估计。CARD\_ 包括在产生的设计中, 是观测量的标识号。

④ 如果实验观测量和模拟观测量重复, 会给出提示报告。

⑤ 如果用户输入的实验观测量(STATUS\_=0)是与 ORTHOPLAN 产生的观测量相同, 只保留一个。

⑥ 偶尔, ORTHOPLAN 会产生两倍的实验观测量。一种处理这些双倍观测量的方法是编辑或删除它们。在这些观测量中设计不再是正交的。另一种选择是可以再运行一次 ORTHOPLAN。当设置不同的种子后再运行一次, ORTHOPLAN 可能产生没有重复观测量的设计。

⑦ ORTHOPLAN 忽略 SPLIT FILE 和 WEIGHT 命令的作用。

## (2) 限制

- ① 不允许有缺失数据
- ② 最多可以指定 10 个因素, 每个因素可以指定 9 个水平。
- ③ ORTHOPLAN 可以产生最多 81 个观测量。

## 4. FACTORS 子命令

FACTORS 子命令指定要在设计中用作因素的变量及其水平值。

(1) 如果数据文件已经存在, 设计产生的观测量附加在数据文件上, 是否使用 FACTOR 子命令是可选的, 如果设计产生的观测量要保存在建立的新数据集或代替当前已经存在的数据文件, 则必须使用 FACTORS 子命令。

(2) 关键字 FACTORS 后面必须跟变量表, 每个变量的标签是可选的, 每个变量的值列表、值标签是可选的。

(3) 值列表和值标签要用括号, 值可以是数值或者可以是在括号中的字符串。

(4) 可选的变量和值标签要加上省略号。

(5) 如果不用 FACTORS 子命令, 在工作数据集中, 除 STATUS\_ 和 CARD 以外的每个变量都被看作是因素变量, 由值标签获得的水平信息在工作数据集定义。ORTHOPLAN 必须在 FACTORS 子命令中或 VALUE LABELS 命令中找到变量值信息。

5. REPLACE 子命令要求用正交设计结果生成或代替当前工作数据集。ORTHOPLAN 可以在数据窗口没有数据时运行, 运行结果生成数据占据数据窗口, 如果数据窗口有工作数据集, 此命令用生成的数据集代替当前工作数据集。

(1) 如果使用了 REPLACE, 那么就要求有 FACTORS 子命令。

(2) 默认运行 ORTHOPLAN 的结果不会代替工作数据集。在 FACTORS 子命令中指定的新变量加上变量 STATUS\_ 和 CARD\_ 附加在工作数据集上。

(3) 当前工作数据集中的数据对要建立的设计文件来说没什么用时, 需要使用 REPLACE。工作数据集将被有 STATUS\_、CARD\_ 变量的和任何其他在 FACTORS 子命令中指定的变量所代替。

6. OUTFILE 子命令把正交设计结果保存到 SPSS 数据文件。

对输出文件只需指定文件名。可以是文件名或以前宣告的数据集的名字。

(1) 默认不创建新数据文件。任何用 FACTORS 指定的新变量加上 STATUS\_ 和 CARD\_ 附加在工作数据文件中。

(2) 输出数据文件包括 STATUS\_、CARD\_ 和所有 FACTORS 子命令中指定的变量。

(3) 由 OUTFILE 产生的文件可以用于其他命令语句, 如 PLANCARDS 和 CONJOINT。

(4) 如果使用了 OUTFILE, 可以不用 REPLACE。

7. MINIMUM 子命令指定最小观测量数。

(1) 不用此命令, 默认产生正交设计必需的最小观测量数。

(2) MINIMUM 后面跟着正整数,这个正整数要小于或等于所有可能的水平组合所能形成的观测量总数。

(3) 如果 ORTHOPLAN 不能产生 MINIMUM 所要求的至少的观测量数,就产生适合指定的因素数和水平数的最大数。

8. HOLDOUT 子命令按关键字后面的数字产生附加在正规设计上的保留观测量。保留观测量由被访者评价,但是在 CONJOINT 估计效应时不用它。

(1) 不指定 HOLDOUT 就不产生保留观测量。

(2) HOLDOUT 后跟正整数,这个正整数要小于等于由所有可能的因素水平组合所形成的观测量总数。

(3) 保留观测量由另一个随机设计产生,不是主效应实验设计。保留观测量不会复制实验观测量,也不会彼此复制。

(4) 实验观测量和保留观测量是在生成的设计中随机混合在一起还是保留观测量附加在实验观测量后面,取决于 MIXHOLD 子命令。保留观测量的 STATUS\_ 变量值是 1。任何模拟观测量都安排在实验观测量和保留观测量后面。

#### 9. MIXHOLD 子命令

MIXHOLD 子命令指定保留观测量是随机地与实验观测量混合还是应该单独出现在文件的实验设计后面。

如果没有指定 MIXHOLD,默认是 NO,意思是在文件中,保留观测量将出现在实验观测量的后面。

(1) MIXHOLD 后面跟着关键字 YES,要求把保留观测量与实验观测量随机混合。

(2) 没有 HOLDOUT 子命令,指定 MIXHOLD 无效。

#### 10. 程序举例

【例2】酸奶的市场调查实验设计程序如下: 见 YUGPLAN1.SPS

```
ORTHOPLAN FACTORS=weight '重量' (600 '600g' 800 '800 g' 1000 '1kg')
```

```
WARRANTY '保质期' (3 '3天' 5 '5 天' 7 '7天') casing '包装' (1 '纸盒' 2 '瓶子')
```

```
/MINIMUM=9 /HOLDOUT=6.
```

##### (1) 程序解释

① ORTHOPLAN 命令后面的 FACTOR 子命令定义了 3 个变量及其水平值和值标签: WEIGHT 变量,标签为“重量”,其值有 3 个水平,以克作单位: 600、800、1000 三个水平值标签分别为 600 克、800 克、1000 克。WARRANTY 变量,标签“保质期”,有 3 个水平值 3、5、7 标签分别为 3 天、5 天、7 天。CASING 变量,变量标签为“包装”,值有两个水平 1、2,分别表示纸盒包装和塑料包装。这些变量及其水平将被用于生成实验设计文件。

② MINIMUM 子命令指定正交实验设计至少要生成 9 个实验观测量; HOLDOUT 子命令指定要生成 6 个保留观测量。

注意,这个程序可以使用中文变量标签和值标签。

## (2) 程序运行结果

数据编辑窗是空窗口，无任何数据时，将产生实验设计数据，包括 5 个变量：WEIGHT、WARRANTY、CASING、STATUS\_和 CARD\_。

另外，还包括 15 个观测量，其中实验观测量 9 个，其 STATUS\_变量值为 0，还有 6 个保留观测量，其 STATUS\_值为 1，排列在 9 个实验观测量后面。见 data16-01。

应该说明的是该程序生成的观测量置于数据窗中，形成当前工作数据集。

如果该程序增加一个子命令：OUTFILE='YUGPLAN.SAV'.则将生成的正交设计保存在命名为YOGPLAN.SAV的数据文件中。

### 【例3】 带有模拟侧面观测量生成的程序

命令语句见文件YUGPLAN2.SPS.内容如下：

```
DATA LIST FREE /WEIGHT WARRANTY CASING.                                ①
VALUE LABELS weight 600 '600g' 800 '800g' 1000 '1kg'                    ②
/WARRANTY 3 '3 days' 5 '5 days' 7 '7days'
/CASING 1 'paper box' 2 'bottle'.
BEGIN DATA                                                                ③
1000 5 1
1000 3 2
END DATA.
ORTHOPLAN.                                                                ④
```

### (1) 程序解释

① DATA LIST命令语句定义了3个变量WEIGHT、WARRANTY、CASING。

② VALUE LABELS命令语句定义了3个变量的水平值和值标签：

WEIGHT变量3个水平，值600、800、1000，标签表明单位为“g”是酸奶的不同重量。

WARRANTY变量3水平，值3、5、7，表明保质期的三个水平为3天、5天、7天。

CASING变量2个水平，值1、2，表示两种包装，纸盒、瓶装。

注意，这里的解释是中文，而程序中变量名、变量标签和值标签都用英文。字符串书写可以是中文。定义变量、变量标签、定义值标签的语句都不可以使用中文。否则显示错误信息。

③ BEGIN DATA –END DATA语句中两行数字是按DATA LIST语句中的变量顺序，给出了2个观测量的三个变量值。这两个观测量在正交实验设计中作为模拟侧面观测量生成的来源。

④ ORTHOPLAN 语句使用上述数据作为正交实验设计的因素、水平和模拟侧面的观测量。无须使用FACTORS子命令定义实验设计需要使用的变量及其水平值。

同样，如果工作数据文件中已经存在设计需要的变量、值及值标签，需要的模拟观测量也已经输入，那么ORTHOPLAN语句无须任何子命令就把数据窗中的所有变量当作

设计需要的因素。

## (2) 运行结果

生成的设计结果见 data16-02.sav。除程序中定义的变量外,还有两个变量。变量 CARDS\_值为观测量号;变量 STATUS\_值表明观测量的性质。有9个观测量的 STATUS\_值为0,它们是实验观测量,2个 STATUS\_值为2是模拟观测量。共11个观测量。

变量有3个,它们的水平数分别是3、3、2个,水平的全组合数是18,输出默认的9种实验组合,即酸奶属性的9个侧面。

数据保存在 data16-02 中。

需要注意的是:如果原数据窗中的数据量很大,变量名与要生成的设计中的变量名还相同,最好将数据窗中的数据清除后再运行新程序。

## 16.3 实验设计结果的打印

在结合分析的研究中,实验设计结果要用来进行调查时显示给被访者,请被访者评分或排序。

SPSS 的正交设计的显示功能可以以两种方式显示或打印正交设计结果:

- ① 粗略的列表格式打印设计结果,以便撰写报告或存档。
- ② 可以显示给被访者观看的,产品的每个属性组合一个个列出的格式。

结果打印程序还允许读者自己加上标题,每个标题占一行;可以打印空行;允许读者加注脚,每个注脚占一行,也可以打印空行。

如果选择列表格式打印,在列表之前打印标题,列表最后打印注脚;

如果选择一个一个列出的格式打印,还可以在每个属性组合之前打印标题,每个属性组合之后打印注脚。

### 16.3.1 设计结果打印过程

从主菜单中逐一选择: Data→Orthogonal Design→Display,打开正交设计结果显示(打印)主对话框,如图16-10所示。操作方法与步骤如下:

(1) 在左面的变量表中选择正交设计的全部因素,将其移到右边的FACTORS栏中。

(2) 在FORMAT 栏选择打印方式:

① Listing for experimenter 对实验侧面使用列表方式显示或打印;分别打印实验侧面、保留侧面,并在它们后面列出模拟侧面。

② Profiles for subjects 打印要呈现在被访者面前的设计侧面,不区分保留侧面、模拟侧面。运行结果是打印全部卡片。

(3) 设置标题和注脚

单击Titles按钮,打开如图16-11的对话框。



① 在Profile Title栏输入标题，也可以输入对被访者的提示，例如排序须知，或者打分方法等说明文字等；

② 在Profile Footer栏输入注脚。

单击 Continue 按钮返回主对话框。

在主对话框中单击OK按钮提交运行。

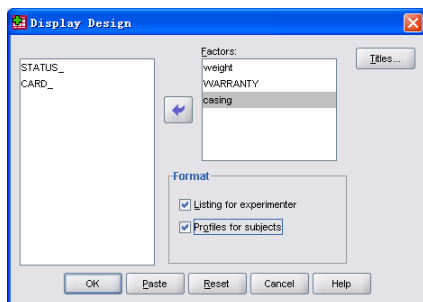


图 16-10 显示功能主对话框

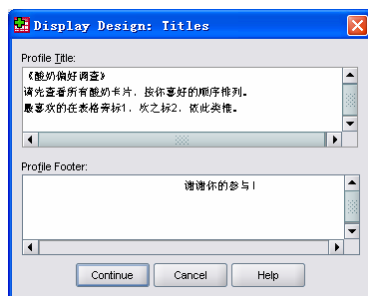


图 16-11 标题、注脚设置对话框

### 16.3.2 打印调查用卡片实例

【例4】以酸奶的偏好调查为例。上一节做好的正交实验设计保存在data16-01中，文件包括实验、保留、模拟三种侧面。

1. 打开文件后的操作如下。

从主菜单中逐一选择：Data→Orthogonal Design→Display，打开正交设计结果显示（打印）主对话框，如图16-10所示。

(1) 在左面的变量表中选择weight、warranty、casing三个因素变量，将其移到右边的FACTORS栏中。

(2) 在FORMAT 栏选择两种打印方式：以列表方式和卡片方式输出。列表方式自己保留设计结果。卡片方式即各侧面一一分别列出，准备呈现在被访者面前使用。

(3) 在Titles对话框中输入标题和注脚。

① 在Titles栏内：

第一行空出，回车后，从第2行开始输入自己打印的标题。因为第一行会有系统默认的标题出现。输入的标题是《酸奶偏好调查》。

第3行输入给被访者的提示：请先查看所有的酸奶卡片，并按您的喜欢程度排序。

第4行输入：最喜欢的请在表格旁标1；次之标2，以此类推。

第5行为空行。

② 在Footer栏输入感谢语：“感谢您的参与！”。

单击Continue按钮返回主对话框。在主对话框中单击OK按钮提交运行。

2. 执行的程序如下:

PLANCARDS

/FACTOR=WEIGHT WARRANTY CASING

/FORMAT BOTH

/TITLE '《酸奶偏好调查》' 请先查看所有酸奶卡片, 按你喜好的顺序排列。最喜欢的在表格旁标 1, 次之标 2, 以此类推。

/FOOTER ' 谢谢你的参与! '

不难读懂这个程序。

PLANCARDS是命令关键字;

FACTORS子命令定义了三个因素变量;

FORMAT子命令选择了两种打印方式;

TITLE子命令在引号中给出标题字符串;

在FOOTER子命令给出的是作为注脚的字符串。

3. 输出结果见图16-4~图16-5。

图16-12是列表式输出, 实验侧面和保留侧面都显示在一个表中。

图16-13没有列出全部卡片。只列出给一个被调查者的第一个实验侧面, 如图16-13(a)所示; 以及第一个保留侧面, 如图16-13(b)所示。

**《酸奶偏好调查》**  
请先查看所有酸奶卡片, 按你喜好的顺序排列。  
最喜欢的在表格旁标1, 次之标2, 依此类推。

	卡标识	WEIGHT	Length of warranty	CASING
1	1	800 g	5days	paperbox
2	2	1kg	5days	paperbox
3	3	1kg	7days	paperbox
4	4	600 g	3 days	paperbox
5	5	600 g	5days	bottle
6	6	800 g	3 days	paperbox
7	7	800 g	7days	bottle
8	8	1kg	3 days	bottle
9	9	600 g	7days	paperbox
10 <sup>a</sup>	10	1kg	3 days	paperbox
11 <sup>a</sup>	11	800 g	5days	bottle
12 <sup>a</sup>	12	600 g	7days	bottle
13 <sup>a</sup>	13	800 g	7days	paperbox
14 <sup>a</sup>	14	800 g	3 days	bottle
15 <sup>a</sup>	15	1kg	5days	bottle

a. 保留

图16-12 各侧面的列表格式输出

**图表文件编号 1: 《酸奶偏好调查》**  
请先查看所有酸奶卡片, 按你喜好的顺序排列。  
最喜欢的在表格旁标1, 次之标2, 依此类推。

卡标识	WEIGHT	Length of warranty	CASING
1	800 g	5days	paperbox

谢谢你的参与!

(a)

**图表文件编号 10: 《酸奶偏好调查》**  
请先查看所有酸奶卡片, 按你喜好的顺序排列。  
最喜欢的在表格旁标1, 次之标2, 依此类推。

卡标识	WEIGHT	Length of warranty	CASING
10	1kg	3 days	paperbox

谢谢你的参与!

(b)

图16-13 实验侧面和保留侧面卡片

### 16.3.3 正交实验设计打印过程语句

PLANCARDS [FACTORS=varlist]

[/FORMAT={LIST}] {CARD} {BOTH}

[/TITLE='string'] [/FOOTER='string'] [/OUTFILE=file]

PLANCARDS为结合分析产生供保留的侧面清单或卡片。设计文件由ORTHOPLAN生成或用户输入。打印的侧面可以用于被访者评价偏爱项目的实验依据。

这个命令读取工作数据集中的数据,以 FACTORS 子命令定义的因素变量为打印的变量,按后续语句要求打印正交实验设计的结果。

过去版本的PAGINATE 子命令已经作废。不能在这里使用。

### 1. 几点说明

除了FACTORS子命令外都是可选的子命令。基本命令就是PALNCARDS。不指定FACTORS,该命令使用当前工作数据集中除STATUS\_和CARD\_以外的所有变量作为打印的基本变量。

① PLANCARD假定工作数据集是结合研究的正交设计结果。在这样的文件中,每一个观测量就是一个结合实验设计的侧面。

② PLANCARD使用在工作数据文件中由ORTHOPLAN产生或由VARIABLE和VALUE LABELS命令产生的因素和因素水平标签。

③ SPLIT FILE 命令对单个卡片输出方式无效。在列表的格式中,每个侧面描述一个不同的设计,并且该命令对每个SPLIT FILE产生的子文件开始一个新的列表。

④ WEIGHT命令对PLANCARD命令无效。

⑤ 不把缺失值当作缺失值,而是当作一个有效值看待。

2. FACTORS 子命令识别要用作因素的变量和它们的标签出现在输出中的顺序。该子命令允许定义字符串变量。

(1) 关键字FACTORS后面跟变量表。

(2) 如果没指定FACTORS,默认工作数据集中除STATUS\_和CARD\_外的所有变量,以它们在文件中出现的顺序作为因素使用。

3. FORMAT子命令指定一个如何显示侧面的方式。选择项是列表方式(关键字LIST)和单个侧面方式(关键字CARD)。

(1) 关键字FORMAT 后面跟着LIST、CARD或BOTH(ALL也可代替BOTH)。

(2) 默认的格式是 LIST。

(3) 用LIST格式,以实验侧面、保留侧面、模拟侧面的顺序列出。用CARD格式,保留侧面作为一个卡片输出,不产生模拟侧面输出。

如果FORMAT=LIST与OUTFILE子命令一起被指定,OUTFILE子命令无效。OUTFILE只对CARD格式有效。与FORMAT=BOTH一起指定OUTFILE时,与OUTFILE、FORMAT=CARD一起使用的效果是相等的。

4. OUTFILE子命令命名一个外部文件,以单个侧面格式写入。列表方式不写到这个外部文件中。

(1) 默认没有到外部文件的输出。

(2) OUTFILE关键字后面跟着一个外部文件,该文件以系统通常的形式指定。

(3) 如果OUTFILE子命令用FORMAT=LIST一起指定,OUTFILE子命令无效。

OUTFILE子命令仅施加于FORMAT=CARD。

5. TITLE子命令指定用于输出的标题,无论是列表格式还是单一侧面格式的顶部的标题字符串。

- (1) 提供默认的标题,除非用OUTFILE子命令直接输出到一个外部文件。
- (2) 关键字TITLE后面跟着用撇号括起来的字符串。
- (3) 如果标题中有撇号,可以用引号代替撇号把字符串括起来。
- (4) 每个TITLE子命令可以指定多个字符串;每个字符串将出现在不同的行中。
- (5) 使用空字符串('')会出现一个空行。
- (6) 可以使用多个TITLE子命令;每个子命令出现在单独的行上。

6. FOOTER子命令指定的字符串将出现在列表格式输出的底部。或者单独侧面格式输出的底部。

- (1) 如果FOOTER在程序中没有使用,在列表和卡片底部是空的。
- (2) FOOTER后面跟着一个放在撇号中的字符串。
- (3) 每个FOOTER子命令可以指定多个字符串。每个字符串出现在一个单独的行上。
- (4) 使用空字符串产生空行。
- (5) 可以指定多个FOOTER子命令,每个子命令会出现在一个单独的行上。

## 16.4 结合分析的语句与编程

打印了供调查使用的卡片,经过培训的调查员就可以按抽样设计矩形调查了。得来的数据经过整理,需要进行结合分析。结合分析的结果提供给决策者作为决策依据。

在 SPSS 的窗口式运行方式的 Analysis 菜单中没有结合分析。要进行结合分析必须进行编程。下面介绍结合分析命令语句及编程要领。

### 16.4.1 结合分析过程语句

1. 结合分析过程 CONJOINT 使用下列命令语句调用

```
CONJOINT [PLAN={* }]{ 'savfile' | 'dataset' }
[/DATA={* }]{ 'savfile' | 'dataset' } / { SEQUENCE } = varlist { RANK } { SCORE }
[/SUBJECT=variable]
[/FACTORS=varlist['labels']
    [{ DISCRETE [ { MORE } ] ] ] { { LESS } } { LINEAR [ { MORE } ] }
    { { LESS } } { IDEAL } { ANTIIDEAL } [ values['labels']] ] varlist...
[/PRINT={ ALL ** } { SUMMARYONLY } ] { ANALYSIS } { SIMULATION } { NONE }
[/UTILITY=file]
[/PLOT={ [ SUMMARY ] [ SUBJECT ] [ ALL ] } ] { [ NONE ** ] }
```

CONJOINT分析偏爱分数或秩数据。由ORTHOPLAN产生的或由用户输入的设计文件描述了在偏爱项目的研究中被打分或排秩的全概念数据集。一种连续或离散模型用于估计每个被访者的或一组被访者的效应。

可以指定怎样把被期望的因素与分数或秩联系起来。

输出可以包括实验数据或模拟数据的分析，或两者都包括。

对每个被访者的效应估计和有关的统计量可以输出到SPSS外部数据文件，以供进一步分析或作图。

## 2. 基本规范

以下基本规范会涉及一些语句。在讲述有关语句的要求和使用方法时不再重复。

(1) 要求有CONJOINT、PLAN或DATA子命令和SEQUENCE、RANK或SCORE子命令描述数据类型。

(2) CONJOINT要求必须有两个文件：设计文件和数据文件。PLAN子命令指定设计文件DATA子命令指定调查数据文件。不一定同时使用两个子命令，可以用PLAN 或DATA子命令中指定一个文件，而当前的工作数据集作为另一个文件。

(3) 默认，由使用DISCRETE模型对设计文件中（除了名为STATUS\_ and CARD的变量）的所有变量计算估计效应。输出包括Kendall's tau和皮尔逊积矩相关系数，预测分数和实际分数之间的相关。显示单尾检验的显著性水平。

(4) 子命令可以是任意顺序。

(5) 可以执行多个FACTORS子命令。而其他子命令，如果一个程序中出现多个，只有最后一个可以执行。

(6) 设计文件和数据文件都可以是外部SPSS数据文件。

在设计文件中，实验侧面的变量STATUS\_必须为0，保留侧面STATUS\_必须为1，模拟侧面的STATUS\_值必须为2。保留侧面由被访者评价，但CONJOINT估计效应时不用。而是在检验效应合法性时使用。模拟侧面是没有被被访者评价的，但是模拟侧面的因素水平组合由CONJOINT根据实验侧面的评价估计其效应。如果没有STATUS\_变量，设计文件中的所有侧面都被假定为实验侧面。

(7) 设计文件中的所有变量除了STATUS\_ 和CARD\_都被CONJOINT作为因素使用。

(8) 除了对每个被访者进行估计以外，对每个在数据文件中定义的分开的文件组计算平均效应。

(9) CONJOINT 检验因素的正交性。如果所有因素都不正交，显示 Cramér 的 V 矩阵统计量，描述非正交性。

(10) 在使用 SEQUENCE 或 RANK 数据时，CONJOINT 对秩尺度进行中心转换，以使计算的系数为正。

(11) 设计文件在收集数据以后不能排序或以任何方法修改，因为设计文件中的侧面顺序必须与数据文件中的数值顺序一一对应（CONJOINT使用的侧面顺序要与它们在设

计文件中出现的顺序一致),不是CARD\_的值决定侧面顺序。如果数据记录方法是RANK或 SCORE,数据文件中第一个被访者的第一个回答就是设计文件中第一个侧面的秩或分数。如果SEQUENCE是数据记录方法,数据文件中第一个被访者的第一个回答是最偏爱的侧面的侧面号(由设计文件中的侧面顺序决定)。

### 3. 限制

(1) 因素必须是数值型变量

(2) 设计文件不能包括缺失值或观测量的权重。在工作数据集中, SUBJECT变量带有缺失值的侧面被聚在一起并在最后计算平均值。如果有被访者的任何一个偏爱数据(秩、分数或侧面号)是缺失的,那个被访者的数据就被跳过,不参与分析。

(3) 因素必须至少有2个水平。每个因素水平最大数为99。

4. PLAN 子命令识别包括全概念侧面的文件。

(1) PLAN子命令关键字后面跟着引用的SPSS数据文件名或当前打开的包括设计的数据集的文件说明。星号代表工作数据集是设计文件。

(2) 程序中如果没有PLAN子命令,当前工作数据文件被认为是默认的设计文件。工作数据文件不能再使用DATA或PLAN子命令指定为设计文件或数据文件。

(3) 设计文件可以由ORTHOPLAN产生,也可由用户直接输入。设计文件可以包括CARD\_ 和 STATUS\_变量,并且必须包括结合分析研究的因素。

5. DATA 子命令指定包括被访者的偏爱分数或秩的(调查数据)文件。

(1) DATA子命令关键字后面跟着被指定的SPSS数据文件或用星号指定当前在数据窗口打开的包括调查数据的数据集文件。

(2) 如果程序中没有DATA子命令,当前工作数据集就是默认的调查数据文件。工作数据文件不能再使用DATA或PLAN子命令指定为设计文件或数据文件。

(3) 在数据文件中的一个变量可以是被访者的标识变量。所有其他变量是被访者的回答数据。并在数量上等于设计文件中的实验侧面和保留侧面的总数。

(4) 被访者的回答可以以秩的形式赋予安排好的侧面顺序,或以分数赋予安排好的侧面顺序,或者侧面号按从最喜欢到最不喜欢的顺序安排。

(5) 允许秩或分数存在结(秩或分数相同的观测量)。如果出现了秩结, CONJOINT发布警告信息然后继续分析。数据以SEQUENCE顺序格式记录时不能有结,因为每个侧面号必须是唯一的。

6. SEQUENCE、RANK、SCORE子命令指定偏爱数据记录的方法。

(1) 必须从三个子命令中选择指定一个,而且只有一个。

(2) 每个子命令后面列出包含偏爱数据的变量名(侧面号变量,秩或分数变量)在设计文件中有多少实验侧面和保留侧面,就必须列出多少变量名。

(3) 子命令关键字含义与规定

① SEQUENCE数据文件中的每个数据点是一个侧面号,以最偏爱的侧面开始并以最

不偏爱的侧面结束。如果被访者被问及从最喜欢到最不喜欢排列侧面卡片。研究人员记录哪个侧面号是第一个, 哪个侧面号是第二个, 等等。

② **RANK** 每个数据点是秩, 从侧面1的秩开始, 然后是侧面2的秩, 以此类推。这就是被访者被要求对每个侧面安排一个秩(顺序), 秩从1到 $n$ , 这里的 $n$ 是侧面数。较低的秩意味较高的偏爱。

③ **SCORE** 每个数据点是赋予该侧面的偏爱分数, 以侧面1分数开始, 然后是侧面2分数, 以此类推。例如通过要求被访者给出从1到100的值表明他们有多喜欢这个侧面。高分对应高偏爱。就是这样的数据类型。

7. **SUBJECT**子命令指定一个标识变量。所有这个变量具有相同值的观测量被组合以便估计效应。

(1) 如果没有使用**SUBJECT**, 所有数据都被假设来自一个被访者, 输出并仅显示一组摘要。

(2) **SUBJECT** 后面跟着变量名。这个变量值标识被访者, 或者标识一组被访对象。

8. **FACTORS**子命令指定要分析的每个因素与秩或分数相关的类型。

(1) 如果没有使用**FACTOR**子命令, 则对所有因素假设为离散模型。

(2) 在设计文件中的所有变量, 除了**CARD\_** 和 **STATUS\_**外都被用作因素, 即使它们没有在**FACTOR**子命令中出现。

(3) **FACTOR**后面跟着变量列表、模型和括号中的模型说明, 该说明描述在秩或分数与变量列表的因素水平之间的期望关系。

(4) 模型说明由模型名和与指定模型的选择项组成。对**DISCRETE**、**LINEAR**模型的选择项**MORE**或**LESS**关键字表明所期望关系的趋势。还可以指定值和值标签。

(5) **MORE**和**LESS**关键字对估计效应不起作用。它们被简单地用作识别那些估计与期望的趋势不一致的观测量(被访者)。四个可用的模型如下:

① **DISCRETE** 因素水平是分类的, 不做因素和分数或秩之间关系的假设。这个设置是默认的。在**DISCRETE**后面指定关键字**MORE**表明因素的高水平被期望是更偏爱; 关键字**LESS**, 表明因素的较低水平被期望是更偏爱。

② **LINEAR** 线性关系。假设期望分数或秩与因素水平的关系是线性的。在**LINEAR**后面指定关键字**MORE**表明因素的高水平被期望是更偏爱。关键字**LESS**表明因素的较低水平期望是更偏爱。

③ **IDEAL** 二次关系描述渐减的偏爱。假设期望分数或秩与因素水平之间是二次关系。它假设存在一个理想的因素水平。与这个理想点的距离, 在任意一个方向上都是与渐减的偏爱相联系。用这个模型描述的因素应该至少有3个水平。

④ **ANTI IDEAL** 二次关系描述渐增的偏爱。期望分数或秩与因素之间的关系是二次的。它假设存在一个最差的因素水平。与这个最差点的距离, 在任意一个方向上都是与渐增的偏爱相联系。用这个模型描述的因素应该至少有3个水平。

(6) 对那些没有列在FACTOR子命令中的变量, 都假设DISCRETE离散模型, 即无模型假设。

(7) 当MORE或LESS关键字与DISCRETE 或 LINEAR一起使用时, 如果所期望的趋势不出现(发生), 会给出注释。

(8) IDEAL 和 ANTIIDEAL两者都生成因素的二次方程。唯一的差别是从与特定点出发, 偏爱还是增加还是减少。对这两个模型的效应估计都相同。当所期望的模型不存在时, 会给出提示。

(9) 选择的值和值标签列表允许记录数据和(或)修改值标签。新值以它们出现在值列表中的顺序以最小的现有值开始替换已经存在的值。如果新值没有指定给一个已经存在的值, 该值保持不变。

(10) 新值标签在撇号或引号中指定。没有新标签的新值保持现有的标签; 新值标签按照它们出现的顺序赋予新值; 如果没有新值赋予它, 以最小的存在的值开始。

(11) 对每个记录的因素, 显示一个标签, 显示原始的记录值和值标签。

(12) 如果因素水平是离散的分类代码(例如1、2、3) 这些值就是CONJOINT在计算中使用的值, 即使值标签包含实际值(例如600、800、1000)。但值标签不会用于计算。你可以用如上所述的值重新编码, 改变代码为实际值。重新编码不会影响DISCRETE因素但是改变了LINEAR、IDEAL和ANTIIDEAL因素的系数。

(13) 对变量的描述输出顺序是所有的DISCRETE变量、LINEAR变量、IDEAL和ANTIIDEAL因素出现在FACTOR子命令中的顺序。

9. PRINT子命令控制输出的内容。输出是否包括对实验数据、对模拟数据的分析结果, 还是两种都包括, 还是没有输出。下列关键字可以使用:

(1) ANALYSIS 输出仅包括实验数据分析的结果。

(2) SIMULATION 仅输出模拟数据的分析结果。三个模拟模型是: 最大效应模型、Bradley-Terry-Luce 即BLT模型和对数模型。

(3) SUMMARYONLY 输出仅包括综合性的概述。这样, 如果被访者很多, 就可以看到综合概述。没必要对每一个被访者都有输出, 使输出量很大。

(4) ALL 输出包括实验数据和模拟数据的分析。ALL是默认的。

(5) NONE 没有输出内容。当你仅想把分析结果写到效应文件时, 采用该关键字。

10. UTILITY 子命令把效应分析结果写到指定的 SPSS 数据文件中。

(1) 程序中没有 UTILITY 子命令, 就没有效应文件输出。

(2) UTILITY 后面跟着要输出的效应文件名。

(3) 该文件使用你的操作系统惯用的指定方法指定。

(4) 效应文件对每个被访者有一个观测量。如果没有使用 SUBJECT, 效应文件包括一个单独的观测量的统计量, 把这组观测量作为一个整体。

(5) 写到效应文件中的变量按下列顺序安排:



- ① 任意一个工作数据集中的SPLIT FILE 变量。
- ② 任意一个SUBJECT变量。
- ③ 回归方程的常数项。对应的变量名为CONSTANT。
- ④ 对DISCRETE 因素, 对被访者估计所有效应。对所有DISCRETE因素估计的效应变量名为因素名后面跟着数字构成。第一个效应后面是1, 第二个效应后面为2, 以此类推。
- ⑤ 对LINEAR因素, 给出单个系数。LINEAR因素的效应名是因素名后面跟着\_L (预测分数的计算是因素值乘系数)。

⑥ 对IDEAL或ANTIIDEAL因素, 因为是二次模型, 所以给出两个系数。系数变量名的命名是在因素名后面分别加\_L (一次项系数) 和\_Q (二次项系数) 构成, (要使用这些系数计算预测分数, 因素值乘以第一个系数加上第二个系数与因素值的平方的乘积)。

⑦ 对设计文件中的所有侧面估计秩或分数。估计的秩或分数的名字对实验和保留侧面用SCORE<sub>n</sub>, 对模拟侧面是SIMUL<sub>n</sub>, 这里的<sub>n</sub>是在设计文件中的位置顺序。即使数据是秩, 实验和保留侧面的名字也是SCORE。

如果生成的变量名太长, 在添加新后缀之前, 从原始变量名结尾截掉字母。

#### 11. PLOT子命令

除了CONJOINT产生的输出外还生成图形。下面是可以用做子命令参数的关键字:

(1) SUMMARY对所有变量产生重要性价值条形图, 和每个变量的效应条形图。如果使用PLOT子命令, 没有给出关键字, 该设置是默认的。

(2) SUBJECT 对每个因素的重要性价值绘制一簇条形图由被访者构成簇, 每个因素一簇条形图, 表明每个因素水平、每个被访者的效应。如果没有SUBJECT子命令指定变量的命名, 就不产生图形, 并显示警告信息。

(3) ALL绘制SUMMARY和SUBJECT两种图。

(4) NONE 不产生任何图形。如果该子命令被省略, 该设置是默认的。

### 16.4.2 结合分析语句实例

**【例5】** 一个结合分析程序的基本结构。

```
CONJOINT PLAN='/DATA/CARPLAN.SAV'                                ①
/FACTORS=WEIGHT (LINEAR MORE) WARRANTY (DISCRETE MORE)
          PRICE (LINEAR LESS)                                       ②
/SUBJECT=SUBJ                                                       ③
/RANK=RANK1 TO RANK15                                              ④
/UTILITY='UTIL.SAV'.                                              ⑤
```

① PLAN子命令指定SPSS外部数据文件CARPLAN.SAV作为设计文件, 它包括全概念侧面。因为没有DATA子命令, 工作数据文件就被假定包括这些侧面被访者评价的数据。

② FACTORS子命令指定因素被期望与秩联系的方法。例如重量被期望与秩线性相

关,所以有较高重量的酸奶将得到较低的秩(更被偏爱,更首选)。WARRANTY因素表明保质期,程序假设为离散型, MORE 表明保质期越长,数值越大,偏爱程度越高。而价格变量即PRICE因素与秩之间是线性关系, LESS表明因素值越高,偏爱程度越低。

③ SUBJECT子命令指定数据集中的SUBJ变量作为标识变量。所有这个变量值相同的观测量被认为是一个观测量,程序将它们组合在一起进行效应估计。

④ RANK子命令指定每个数据点是特定侧面的秩,共有15个变量对应这些秩值。而且在包括这些秩的工作数据集中识别这些变量。

⑤ UTILITY把输出结果写入一个外部数据文件,名为 UTIL.SAV,它包括每个被访者的效应估计和有关的统计量。

以上程序表明一个基本的结合分析程序应该明确,设计文件、调查数据文件的位置或名称。在调查数据文件中一定要有观测量的标识变量,否则就会把所有观测量当成一个被访者的数据进行处理。

使用FACTORS因素指定因素变量虽然可以省略,省略的结果,程序会把设计文件中除STATUS\_、CARD\_变量外都作为因素变量,但是要探讨秩或分数与因素之间的关系时还是不可没有FACTORS子命令的。

【例6】 本例仍然使用酸奶偏爱研究的思路。应该主要注意与CONJOINT命令有关的数据来源与文件指定。

```
DATA LIST FREE /CARD_  WARRANTY  WEIGHT  CASING  STATUS_  ①
BEGIN DATA  ②
1 5 1 600 2
2 5 2 600 2
3 3 2 800 2
4 3 1 1000 2
END DATA. ③
ADD FILES FILE='/DATA/YUGPLAN.SAV'/FILE=*. ④
CONJOINT PLAN=* ⑤
/DATA='/DATA/YUGDATA.SAV' ⑥
/FACTORS= WEIGHT (LINEAR) WARRANTY (DISCRETE MORE) ⑦
/SUBJECT=SUBJ /RANK=RANK1 TO RANK15 ⑧
/PRINT=SIMULATION. ⑨
```

① DATA LIST定义5个变量, CARD\_标识变量, 3个因素变量和STATUS\_变量。

②~③ BEGIN DATA和END DATA之间的数据是4个模拟侧面。每个侧面包括一个CARD\_标识号和感兴趣的因素水平的特殊组合。这4个侧面被生成在数据窗中,成为工作数据文件。

所有侧面(观测量)的STATUS\_变量值都等于2。CONJOINT认为这些STATUS\_=2

的是模拟侧面。

④ ADD FILES命令是合并数据文件的过程命令语句。命令后面必须跟“FILE=”指定一个数据文件YUGPLAN.SAV, 用FIL=\*子命令指定工作数据文件。注意工作数据文件在ADD FILES命令最后指定, 所以模拟侧面数据是附加在YUGPLAN.SAV的末尾构成新的工作数据集。

⑤ CONJOINT中的 PLAN子命令定义这个新的工作数据集作为设计文件。

⑥ DATA子命令指定一个CONJOINT要分析的数据文件YUGDATA.SAV。

⑦ FACTORS子命令指定因素WEIGHT重量与秩数据预期是线性关系, 而保质期WO-RRANTY是离散数据, 但与秩数据的预期关系是保质期越长, 数值越大, 偏爱程度越高。

⑧ SUBJECT子命令和RANK子命令的语句形式和功能与例1相同。

⑨ PRINT子命令指定只输出模拟观测量的分析结果。

【例7】这是一个有关汽车的偏爱分析研究的例题。因素有WARRANTY保质期(1、3、5年, 3个水平)、SEATS座位数(2、4座, 2个水平)、PRICE价格(7000、10000、14000, 单位美元, 3个水平) SPPED最高速度(70、100、130英里/小时)。被访者对设计文件中的15个侧面排秩, 显然, 秩值自1~15。下面的程序, 秩数据由程序输入运行保存到数据文件中, 设计的15个侧面数据也由程序输入, 运行, 存在工作数据文件中。

```
DATA LIST FREE /SUBJ RANK1 TO RANK15.
```

①

```
BEGIN DATA
```

②

```
01 3 7 6 1 2 4 9 12 15 13 14 5 8 10 11
```

```
02 7 3 4 9 6 15 10 13 5 11 1 8 4 2 12
```

```
03 12 13 5 1 14 8 11 2 7 6 3 4 15 9 10
```

```
04 3 6 7 4 2 1 9 12 15 11 14 5 8 10 13
```

```
05 9 3 4 7 6 10 15 13 5 12 1 8 4 2 11
```

```
50 12 13 8 1 14 5 11 6 7 2 3 4 15 10 9
```

```
END DATA.
```

③

```
SAVE OUTFILE= '/DATA/RANKINGS.SAV'.
```

④

```
DATA LIST FREE /CARD_ WARRANTY SEATS PRICE SPPED .
```

⑤

```
BEGIN DATA
```

⑥

```
1 1 4 14000 130
```

```
2 1 4 14000 100
```

```
3 3 4 14000 130
```

```
4 3 4 14000 100
```

```
5 5 2 10000 130
```

```
6 1 4 10000 070
```

```

7 3 4 10000 070
8 5 2 10000 100
9 1 4 07000 130
10 1 4 07000 100
11 5 2 07000 070
12 5 4 07000 070
13 1 4 07000 070
14 5 2 10000 070
15 5 2 14000 130
END DATA. ⑦
CONJOINT PLAN=* /DATA=' RANKINGS.SAV' ⑧
/FACTORS=PRICE (ANTIIDEAL)SPEED (LINEAR)
WARRANTY (DISCRETE MORE) ⑨
/SUBJECT=SUBJ /RANK=RANK1 TO RANK15. ⑩

```

① 第一个DATA LIST 定义了15个变量，第一个是观测量号变量SUBJ，其后的15个变量是被访者评价的秩，命名为RANK1~RANK15。

②③ BEGIN-END组， DATA命令产生包含秩的数据文件。

④ 由BEGIN-END组的数据产生的文件保存在外部文件RANKINGS.SAV中。

⑤ 第二个DATA LIST 定义的是正交设计中的4个因素变量。

⑥⑦ 第2个BEGIN-END DATA组，共15个侧面数据。没有保存语句跟在后面，所有运行后在数据窗中，即作为工作数据文件。

⑧ CONJOINT命令中PLAN=\* 使用工作数据文件作为设计文件，DATA子命令指定了外部数据文件RANKINGS.SAV 作为数据文件。

⑨ FACTORS子命令，假设价格因素PRICE水平与秩值是倒二次的关系，存在一个被访者认为最差的价格水平，其他价格水平的偏爱都高于此水平；假设速度因素SPEED水平值与秩值之间是线性关系；假设保质期是离散的，保质期越高，秩值越高即更加偏爱。

⑩ SUBJECT子命令定义在数据文件中，被访者识别号变量为SUBJ；RANK子命令定义秩值变量是RANK1~RANK15。

上述程序的最后三行，⑧⑨⑩，可以写成：

```

CONJOINT PLAN=* /DATA=' RANKINGS.SAV'
/FACTORS=PRICE (ANTIIDEAL) WEIGHT (LINEAR) WARRANTY (DISCRETE
MORE)
/SUBJECT=SUBJ /RANK=RANK1 TO RANK15.

```

需要说明的是：

① RANK子命令指定的数据是按排号顺序安排的侧面的秩。在SUBJ后面第一个数据

点是变量RANK1，它是第一个被访者给第一个侧面的秩。

② 在设计文件中有15个侧面，所以必须有15个秩变量。

③ 该例题使用了TO关键字指示有15个秩变量。

【例8】 仍然是汽车的偏爱研究，主要看FACTORS子命令的赋值功能。

```
CONJOINT DATA='DATA.SAV'                                ①
/FACTORS=PRICE (LINEAR LESS) WEIGHT (IDEAL 70 100 130)
WARRANTY (DISCRETE MORE)                                ②
/SUBJECT=NO                                              ③
/RANK=RANK1 TO RANK15.                                  ④
```

① CONJOINT命令使用DATA指定数据文件。它至少应该包括16个变量。除了RANK1~RANK15外的变量应该是③中SUBJECT子命令指定的变量NO是观测号。

② FACTOR 子命令指定期望相关。期望价格和秩之间是线性相关，所以较高的价格偏爱较低（高秩）。期望在速度水平与秩之间是二次相关，期望较长的保证期与较大的偏爱（低秩）对应。

③ WEIGHT因素有一个新值列表。如果原来的值为代码1、2、3，那么70代替1，100代替2，130代替3。

任何设计文件中没有列在FACTOR子命令中的变量除了CARD\_和STATUS\_外，都用DISCRETE模型。

# 16.5 结合分析实例

## 16.5.1 课题分析与正交设计

【例 9】 这里要研究地毯吸尘器的顾客偏爱的课题。

### 1. 课题内容与因素、因素水平的选择

这是一个流行的结合分析的例题。(Green and Wind, 1973)一个公司对地毯吸尘器的销售感兴趣，希望调查 5 个对消费者偏爱的影响因素，包装设计、商标名称、价格、好管家封条、货币的售后保证。经过认真考虑，包装设计有 3 个水平，每个水平表明刷子的不同位置；商标名字三个水平（K2R, Glory, and Bissell）；三种价格水平；好管家封条两个水平（有、否）；售后的现金保障，两个水平（是、否）。表 16-4 显示了在地毯吸尘器研究中的变量名与变量标签。表 16-5 是各变量的值和值标签。

或许还有其他因素和因素水平描述地毯清洁器，但是对管理者来说只有这些是感兴趣的。对结合分析来说这一点是很重要的。要选择的参与研究的因素必须是你认为最影响偏爱的因素变量。使用结合分析，将开发出基于这 5 个因素的顾客偏爱模型。

结合分析的第一步是产生一个展现在被访者面前的产品侧面的因素水平组合。由于即使很少的因素数和每个因素很少的水平数也会导致一个处理不了的产品侧面的数量。

因此需要产生典型的子集，即正交设计的安排。

表 16-4 变量表

变量名	变量标签
Package	包装设计
brand	商标名称
price	价格
seal	好管家封条
money	货币式售后保证

表 16-5 值和值标签对应表

因素	值	值标签
package	1, 2, 3	A*, B*, C*
brand	1, 2, 3	K2R, Glory, Bissell
price	1.19, 1.39, 1.59	\$1.19, \$1.39, \$1.59
seal	1, 2	no, yes
money	1, 2	no, yes

Generate Orthogonal Design 程序产生正交安排还涉及正交设计及保存信息到 SPSS 文件中。不像大多数程序那样，而是在运行 Generate Orthogonal Design 之前不是必须有一个工作数据集。如果没有工作数据集，则可以选择生成一个，产生变量名、变量标签、和在对话框选择项中选择值标签。如果已经有了工作数据集，则可以代替它，或者保存正交设计作为一个 SPSS 数据文件。

下面的操作产生正交设计数据。在操作之前数据窗口是空的。

2. 正交设计操作过程

(1) 按 Data→Orthogonal Design→Generate 顺序单击菜单项见图 16-4，打开 Generate Orthogonal Design 对话框，见图 16-5。

(2) 定义因素变量名及其标签

将表 16-6 中的第 1 个变量名 Package 输入 Factor Name 栏，将其变量标签包装设计输入 Factor Label 栏，单击 Add 按钮，送入 Add 按钮旁的矩形框内。在矩形框内显示 Packege'包装设计' (?)。

再将表 16-6 中第 2 个变量名 Brand 和标签‘商标名称’分别输入到 Factor Name 栏和 Factor Label 栏，单击 Add 按钮，送入 Add 按钮旁的矩形框内。在矩形框内显示 Brand '商标名称' (?)。以此类推。具体操作参考 16.2.3 节中的说明。

(3) 定义各因素变量的值和值标签

以定义第一个变量的值和值标签为例说明操作。

在矩形框中选择（单击）第一项 Packege‘包装设计’(?)矩形框下面的 Define Value 按钮，打开 Orthogonal Design-Define Value 对话框。在 Value 列的 1、2、3 行分别输入 1、2、3，在 Label 列的 1、2、3 行对应位置分别输入值标签 A\*、B\*、C\*。单击 Continue 按钮，返回主对话框。

在主对话框中，在第一个变量定义的位置显示 Package'包装设计' (1'A\* 2'B\* 3'C\*)。这样一个个定义变量。

(4) 在主对话框中选择 Data File 栏中的第 2 项: Replace working data file, 要求生成的设计代替当前的工作数据文件。即生成在数据窗中。

(5) 指定随机数种子。在 Reset random number seed to 后面输入随机数种子 2000000。

(6) 指定生成设计的观测量数。单击主对话框中的 Options 按钮打开如图 16-7 的对话框。

① 在 Minimum number of cases to generate 栏输入 18; 最小观测量数是 16, 而根据

需要, 还要求多 2 个侧面, 所以输入 18。要求产生 18 种水平组合的侧面数据。

② 在 Hold cases 栏中选择 Number of holdout cases, 并在其后输入保留观测量数 4。

单击 Continue 按钮, 返回主对话框。在主对话框中单击 OK 按钮, 提交系统执行。

在数据窗中生成设计结果, 见图 16-14 中第 1~18 个观测量。

图 16-14 中变量 STATUS\_ 值为 0 的是实验侧面, 值为 1 的是保留侧面。在原设计结果中再输入两个模拟观测量, 其 STATUS\_ 变量的值是 2。这个数据集保存为 data16-03.sav。

(7) 生成设计文件后, 还应该对所生成的保留侧面和模拟侧面进行查重。如果保留侧面与实验侧面有重复, 再重复上述操作, 直到保留侧面与设计侧面没有重复为止。可

	package	brand	price	seal	money	STATUS_	CARD_
1	1.00	2.00	1.39	2.00	1.00	0	1
2	2.00	1.00	1.19	1.00	1.00	0	2
3	2.00	2.00	1.39	1.00	2.00	0	3
4	3.00	2.00	1.59	1.00	1.00	0	4
5	3.00	3.00	1.39	1.00	1.00	0	5
6	1.00	3.00	1.39	1.00	1.00	0	6
7	2.00	3.00	1.59	2.00	1.00	0	7
8	1.00	1.00	1.59	1.00	2.00	0	8
9	3.00	1.00	1.39	1.00	1.00	0	9
10	3.00	2.00	1.19	1.00	2.00	0	10
11	3.00	1.00	1.59	2.00	1.00	0	11
12	2.00	2.00	1.59	1.00	1.00	0	12
13	3.00	3.00	1.19	2.00	2.00	0	13
14	1.00	2.00	1.19	2.00	1.00	0	14
15	2.00	1.00	1.39	2.00	2.00	0	15
16	1.00	1.00	1.19	1.00	1.00	0	16
17	1.00	3.00	1.59	1.00	2.00	0	17
18	2.00	3.00	1.19	1.00	1.00	0	18
19	1.00	3.00	1.59	2.00	1.00	1	19
20	3.00	1.00	1.19	2.00	1.00	1	20
21	1.00	2.00	1.59	1.00	1.00	1	21
22	1.00	3.00	1.19	1.00	1.00	1	22

图 16-14 数据窗中的设计结果

可以说这种保留侧面与实验侧面重复的现象是偶尔出现的, 但是每次实验设计结束后都要查重, 保证设计无误。查重使用 DATA 菜单的 Identify Duplicate cases 功能, 详见第 2 章有关内容。

## 16.5.2 调查准备与调查

1. 将设计文件打印成调查用的卡片和存档用的文件。

(1) 从主菜单中逐一选择: Data→Orthogonal Design→Display, 打开正交设计结果显示 (打印) 主对话框, 见图 16-10。

(2) 在左面的变量表中选择 Package、Brand、Price、Seal、Money 五个因素变量, 将其移到右边的 FACTORS 栏中。

(3) 在 FORMAT 栏选择两种打印方式: 以列表方式和卡片方式输出。

(4) 在主对话框中单击 Titles 按钮, 打开如图 16-15 所示的标题、注脚设置对话框。在 Titles 栏内做如下操作:

第一行空出，回车后，因为第一行会有系统默认标题出现。从第 2 行开始输入自己打印的标题。输入的标题是《地毯清洁剂调查》等，见图 16-15。

在 Footer 栏输入提示：请检查所填写的序号是否有重复！。感谢语：谢谢参与！

单击 Continue 按钮返回主对话框。在主对话框中单击 OK 按钮提交运行。

输出结果见图 16-16、图 16-17 和图 16-18。

## 2. 调查

我们将如图 16-17 的整个设计文件列表存档。

是该项研究设计结果，前 18 个观测量是正交设计结果。还有 4 个保留侧面，2 个模拟侧面。

将如图 16-16 的卡片打印多份(每份 22 张卡片)，用于市场调查。让每个被访者将 22 个卡片认真浏览后，按最喜欢到最不喜欢的顺序排序，将最喜欢的标 1；次之的标 2；以此类推，最不喜欢的标 22。

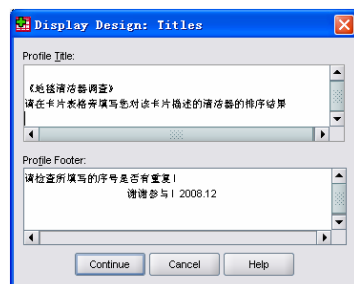


图 16-15 标题与注脚

概要文件编号 1:  
《地毯清洁剂调查》

请在卡片表格旁填写您对该卡片描述的清洁剂的排序结果

卡标识	package design	brand name	price	Good Housekeepin g seal	money-back guarantee
1	A*	Glory	\$1.39	yes	no

请检查所填写的序号是否有重复! 谢谢参与!

(a)

概要文件编号 18:  
《地毯清洁剂调查》

请在卡片表格旁填写您对该卡片描述的清洁剂的排序结果

卡标识	package design	brand name	price	Good Housekeepin g seal	money-back guarantee
18	B*	Bissell	\$1.19	no	no

请检查所填写的序号是否有重复! 谢谢参与!

(b)

概要文件编号 19:  
《地毯清洁剂调查》

请在卡片表格旁填写您对该卡片描述的清洁剂的排序结果

卡标识	package design	brand name	price	Good Housekeepin g seal	money-back guarantee
19	A*	Bissell	\$1.59	yes	no

请检查所填写的序号是否有重复! 谢谢参与!

(c)

图 16-16 实验、保留、模拟侧面卡片（举例）

《地毯清洁剂调查》

请在卡片表格旁填写您对该卡片描述的清洁剂的排序结果

	卡标识	package design	brand name	price	Good Housekeepin g seal	money-back guarantee
1	1	A*	Glory	\$1.39	yes	no
2	2	B*	K2R	\$1.19	no	no
3	3	B*	Glory	\$1.39	no	yes
4	4	C*	Glory	\$1.59	no	no
5	5	C*	Bissell	\$1.39	no	no
6	6	A*	Bissell	\$1.39	no	no
7	7	B*	Bissell	\$1.59	yes	no
8	8	A*	K2R	\$1.59	no	yes
9	9	C*	K2R	\$1.39	no	no
10	10	C*	Glory	\$1.19	no	yes
11	11	C*	K2R	\$1.59	yes	no
12	12	B*	Glory	\$1.59	no	no
13	13	C*	Bissell	\$1.19	yes	yes
14	14	A*	Glory	\$1.19	yes	no
15	15	B*	K2R	\$1.39	yes	yes
16	16	A*	K2R	\$1.19	no	no
17	17	A*	Bissell	\$1.59	no	yes
18	18	B*	Bissell	\$1.19	no	no
19*	19	A*	Bissell	\$1.59	yes	no
20*	20	C*	K2R	\$1.19	yes	no
21*	21	A*	Glory	\$1.59	no	no
22*	22	A*	Bissell	\$1.19	no	no
23*	1	C*	K2R	\$1.19	no	no
24*	2	B*	Glory	\$1.19	yes	yes

请检查所填写的序号是否有重复! 谢谢参与!

a. 保留  
b. 模拟

图 16-17 列表格式输出

得到数据后，输入到数据编辑窗，形成数据文件。见图 16-18。

对每一个被访者建立一个观测量，有一个标识号即顺序号；另外 22 个变量分别是被访者对 22 个侧面的排序结果，由于无重复，可以认为每个卡片上标的都是被访者给出的卡片所示侧面的秩。输入数据后的数据窗见图 16-18。保存在 data16-04 中。



16.5.3 结合分析编程与结果分析

1. 安排数据文件和设计文件

在数据编辑窗中打开 data16-04 作为结合分析的数据文件，见图 16-18。

设计文件在 E:\书\SPSS 统计分析第 4 版\data\文件夹中。名为 data16-03.sav。

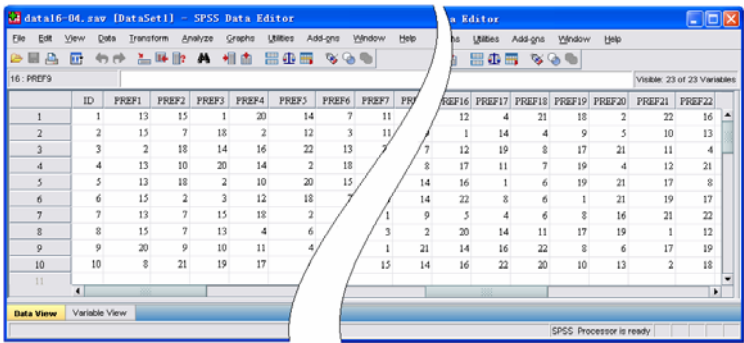


图 16-18 调查数据文件

2. 程序清单如下

```
CONJOINT PLAN='E:\书\SPSS 统计分析第 4 版\data\data16-03.sav'  
  /DATA=*      /SEQUENCE=PREF1 TO PREF22    /SUBJECT=ID  
  /FACTORS=PACKAGE BRAND (DISCRETE) PRICE (LINEAR LESS)  
          SEAL (LINEAR MORE) MONEY (LINEAR MORE)  
  /PRINT=SUMMARYONLY.
```

3. 程序解释

PLAN 子命令指定设计文件的位置和文件名；

注意，读者在运行程序时需要在 PLAN 子命令中给出自己的文件存储路径，程序方能正常执行。

DATA 子命令指定调查数据文件在数据窗中；

SEQUENCE 子命令指定变量 PREF1~PREF22 分别表示最喜欢的侧面号至最不喜欢的侧面号。例如第一个观测量的 PREF1 的值就是第一个被访者最喜欢的侧面号。

SUBJECT 子命令指定观测量的标识变量为 ID。

FACTORS 子命令指定 5 个因素变量，是设计文件中除 CARD\_、STATUS\_外的所有变量；指定这些变量与秩的预期关系：价格变量 PRICE、好管家封条变量 SEAL、货币售后保证 MONEY 与秩预期均为线性关系，LESS 表明预期的价格越低秩越低偏爱程度越高，封条是代码，1 是 no，2 代表 Yes，SEAL 与秩的线性预期参数是 MORE，表明预期被访者更偏爱有好管家封条的产品；同样理解货币售后的保证，预期被访者更偏爱有货币售后保证的产品。

PRINT 子命令指定只打印综合分析表。

4. 运行结果见表 16-6~表 16-15。

5. 结果解释

表 16-6 是正交检验结果，无论检验结果如何都给出警告信息。本例设计经检验为正交设计。

表 16-7 是模型描述。

表 16-8 是整体效应估计。

如所预期的一样，在价格 price 与效应之间存在负的关系。较高的价格与较低的效应相关（大的负值意味着较低的效应）。也正如所预期的，对封条或货币售后保证与较高的效应相应。

表 16-7 模型描述

Model Description		
	N of Levels	Relation to Ranks or Scores
package	3	Discrete
brand	3	Discrete
price	3	Linear (less)
seal	2	Linear (more)
money	2	Linear (more)

表16-6 正交检验结果

Warnings<sup>a</sup>

All factors are orthogonal.

表 16-8 整体效应估计

		Utilities	
		Utility Estimate	Std. Error
package	A*	-2.233	.192
	B*	1.867	.192
	C*	.367	.192
brand	K2R	.367	.192
	Glory	-.350	.192
	Bissell	-.017	.192
price	\$1.19	-6.595	.988
	\$1.39	-7.703	1.154
	\$1.59	-8.811	1.320
seal	no	2.000	.287
	yes	4.000	.575
money	no	1.250	.287
	yes	2.500	.575
(Constant)		12.870	1.282

表 16-9 重要性值

Importance Values	
package	35.635
brand	14.911
price	29.410
seal	11.172
money	8.872
Averaged Importance Score	

表 16-10 回归系数

Coefficients	
	B Coefficient
	Estimate
price	-5.542
seal	2.000
money	1.250

表 16-11 相关性检验

Correlations <sup>a</sup>		
	Value	Sig.
Pearson's R	.982	.00
Kendall's tau	.892	.00
Kendall's tau for Holdouts	.667	.08
a. Correlations between observed and estimated preferences		

表 16-12 模拟侧面的偏好分数

Preference Scores of Simulations <sup>a</sup>		
Case	ID	Score
1	1	10.258
2	2	14.292

a. Negative simulation scores or all zero simulation scores are found. This subject will not be included in computing preference probabilities using the Bradley-Terry-Luce or Logit methods.

表 16-13 模拟侧面的偏好概率

Preference Probabilities of Simulations <sup>a</sup>				
Card No.	ID	Maximum Utility <sup>b</sup>	Bradley-Terry-Luce	Logit
1	1	30.0%	43.1%	30.9%
2	2	70.0%	56.9%	69.1%

a. Including tied simulations

b. y out of x subjects are used in the Bradley-Terry-Luce and Logit methods because these subjects have all nonnegative scores.

表 16-14 逆相关小结

Reversal Summary	
N	N of Subjects
1	3
2	2

This table displays the number of subjects that have the given number of reversals.

由于在表 16-8 中列出了所有因素的各水平的效应。因此可以各因素选择一个水平，代表组合成感兴趣的侧面，而这个侧面不一定是在设计中出现的侧面。在表中查出它们的效应，加在一起给出任何一个组合的总效应。

例如，具有包装设计 B\*、商标 Bissell、价格\$1.59、没有好管家封条和货币式售后保证的清洁器的总效应是：

utility(package B\*) + utility(Bissell) + utility(\$1.59) + utility(no seal) + utility(no money-back) + constant=1.867 + (-0.017) + (-8.811) + 2.000 + 1.250 + 12.870 = 9.159

如果清洁器的包装设计为 C\*，商标为 K2R，价格\$1.39，有认可的好管家封条和货币式售后保证，总效应就是 0.367 + 0.367 + (-7.703) + 4.000 + 2.500 + 12.870 = 12.401

这两种地毯清洁器比较，顾客更偏爱后者。

进一步还可以计算顾客最不喜欢的组合和最偏爱的组合，显然最喜欢的组合是包装为 B\*，商标为 K2R，价格 1.19，有好管家封条和货币式售后保证的。当然对商家来说，还要与利润一同综合考虑。可能选择的商品是在保证利润的前提下，总效应又比较高的。

表 16-9 是相对重要性值。这个表提供了每个因素相对重要性的测度，即重要性分数或重要性值。该值的计算方法是先对每个被访者计算每个因素的效应范围，除以所有因素的效应范围的总和，用百分比表示。再对所有被试者的该因素效应取平均值。重要性值高的因素在顾客看来相对更重要。

表 16-15 逆相关统计

Number of Reversals		
Factor	price	3
	money	2
	seal	2
	brand	0
	package	0
Subject	1 Subject 1	1
	2 Subject 2	2
	3 Subject 3	0
	4 Subject 4	0
	5 Subject 5	0
	6 Subject 6	1
	7 Subject 7	0
	8 Subject 8	0
	9 Subject 9	1
	10 Subject 10	2

对没有 SUBJECT 子命令，不对每个被访者进行计算，是将整个数据文件看作一个被访者计算总效应。重要性计算就像对一个被访者所进行的计算一样。

然而，当使用了 SUBJECT 子命令时，对每一个单独的被访者是被平均的，这些平均的重要性将不会与那些使用总效应的计算相一致。

这个结果表明包装设计对整个偏爱最有影响力。这就意味在产品侧面之间存在大的偏爱差异，包括最小的包装要求。结果还表明货币式售后保证在整个决定偏爱中重要性最小。价格是个有重要意义的角色，但是不如包装设计那样大。或许这是因为价格水平之间的差距不是很大的缘故。

表 16-10 为回归系数。这个表表明 LINEAR 所指定的因素的线性回归系数(程序中没有指定二次模型 IDEAL 和 ANTIIDEAL 模型, 如果指定了, 或许存在二次项)。特定因素水平的效应由水平与系数相乘来确定。例如对价格 Priced\$1.19 的预期效应如效应表中所示的-6.595。这是简单地把价格的水平值 1.19，乘以价格系数-5.542 的结果。

表 16-11 为相关性检验。表格提供了两个统计量，皮尔逊 R 和肯道尔 τ，是观测量的和估计参数之间的相关测度。表格还对保留侧面显示了 Kendall's tau。本例有 4 个保留侧面，是由课题决定的没有被结合分析过程使用来估计效应。而结合分析过程对这些侧面计算观测量的和预测的秩之间的相关是作为对效应有效性的检验。

在许多结合分析中，参数的数量与设计的侧面数关系密切。这会使观测量的和估计的分数之间的相关人为地膨胀（增高）。在这种情况下，保留侧面的相关可能给出比较好的对模型拟合的指示。然而要留神，保留侧面将会产生比较低的相关。

表 16-12 和表 16-15 是对模拟侧面的分析。模拟侧面是人为输入的两个感兴趣的侧面，不是设计自动生成的。所以在计算估计效应时没有使用这两个侧面。这两个侧面是：

① 包装 package 水平为 C\*、商标 Brand 水平为 K2R、价格 Price 水平为\$1.19、封条和货币式售后保证均为 no。

② 包装 package 水平为 B\*、商标 Brand 水平为 Glory、价格 Price 水平为\$1.19、封条和货币式售后保证均为 Yes。

查表 16-8 整体效应估计表中的各因素水平的效应估计值，相加得到模拟观测量①的效应值为： $0.367+0.367-6.595+2+1.25+12.870=10.295$

模拟观测量②的效应值为： $1.867-0.35-6.595+4+2.5+12.870=14.292$

表 16-13 模拟侧面的偏好概率表给出了三种模型预测每个模拟侧面可能成为最偏爱的一种属性组合的可能性。

可以看出任何一种模型预测的结果都是第 2 个模拟侧面的概率大于第一个模拟侧面。因此选择第 2 个模拟侧面所表达的属性组合作为最偏好的属性组合的可能性最大。

表 16-14 为逆相关小结。

在 FACTORS 子命令中给出了对三个因素的预测模型类型：对价格的预测是线性模型，LESS 关键字给出预测方向，即预测被访者对高价格有较低的偏爱；对封条 Seal 和售后 Money 两个因素的预测是线性模型，方向是关键字 MORE，预测对 Yes (2) 比 no (1) 有更高的偏爱。有些被访者给出的秩表明偏爱与所预期的相反，程序就会对这种情况给以记录和统计。该表显示，被访者选择与预期相反的情况发生一次的有 3 个被访者，发生 2 次的有 2 个被访者。

表 16-15 为逆相关统计详表。

表明 3 个被访者的选择对价格 Price 因素不是认为越低越好；2 个被访者的选择，对货币式售后 Money 因素不是认为有比没有更好；2 个被访者对封条 Seal 因素不是认为有比没有更好。由于包装 Package、商标 Brand 在 Factors 子命令中指定为离散因素，所以没有逆相关的问题。统计数自然为 0。

在表中的 Subject 的部分，列出了逆相关发生在哪几个被访者身上，发生了几次。对这几个被访者的回答还可以进行详细研究。

6. 如果程序最后一个语句改变为：

/PRINT=ALL.

数据文件安排与其他语句全部与 5 中所叙述一样。那么输出还会包括对每个被访者数据的一一分析，例如对第 5 个被访者的输出，如表 16-16 所示。

据计算出的各因素的各水平的效应与标准误：

(a) 是根据第 5 个被访者数据计算出的各因素效应估计值；

(b) 是根据第 5 个被访者数据计算出的各因素重要性值；

(c) 是根据第 5 个被访者数据计算出的三个线性模型因素的回归系数；

- (d) 是根据第 5 个被访者数据计算出的观测量的和估计参数之间的相关测度。两个统计量，皮尔逊  $R$  和肯道尔  $\tau$ ；
- (e) 是根据第 5 个被访者数据计算出的两个模拟侧面的偏爱分数。

表 16-16 对每个被访者数据的分析输出

Utilities			
		Utility Estimate	Std. Error
package	A*	-6.000	.313
	B*	3.000	.313
	C*	3.000	.313
brand	K2R	.167	.313
	Glory	-1.000	.313
	Bissell	.833	.313
price	\$1.19	-19.833	1.614
	\$1.39	-23.167	1.886
	\$1.59	-26.500	2.157
seal	no	1.000	.470
	yes	2.000	.940
money	no	1.000	.470
	yes	2.000	.940
(Constant)		30.000	2.095

(a)

Importance Values	
package	46.154
brand	9.402
price	34.188
seal	5.128
money	5.128

(b)

Coefficients		
	B Coefficient	
	Estimate	Std. Error
price	-16.667	1.357
seal	1.000	.470
money	1.000	.470

(c)

Correlations <sup>a</sup>		
	Value	Sig.
Pearson's R	.991	.000
Kendall's tau	.957	.000
Kendall's tau for Holdouts	1.000	.021

a. Correlations between observed and estimated preferences

(d)

Preference Scores of Simulations <sup>a</sup>		
Case	ID	Score
1	1	15.333
2	2	16.167

a. Negative simulation scores or all zero simulation scores are found. This subject will not be included in computing preference probabilities using the Bradley-Terry-Luce or Logit methods.

(e)

习 题 16

1. 对于市场调查中顾客偏爱的分析必须要用结合分析吗？
2. 结合分析适用于什么样的数据，主要解决什么问题？
3. 要调查分析某产品不同侧面组合的顾客偏爱，整个工作要分哪几个主要步骤，每个步骤可以用 SPSS 的哪些程序解决，每个步骤的作用是什么？
4. 用 Conjoint 命令语句编程，必须包括什么语句？
5. 如果数据窗中有实验设计数据，那么程序中可以减少哪个语句？
6. 市场上先调查了解市民层购买的酸奶有几种，主要的重量等级、品牌、价格、保质期。每个因素取 2~3 个等级，设计使用结合分析了解市民偏爱的课题解决方案。如果可能，将调查数据使用程序分析并得出结论。
7. 设计一个台式个人计算机的顾客偏爱课题及其解决方案。

## 第 17 章 时间序列分析

时间序列是指依时间顺序取得的观察资料的集合。在一个时间序列中，离散样本序列可以按相等时间间隔或不相等时间间隔获取，更多的是采用前者来实现。时间序列的特点是数据资料的先后顺序不能随意地改变，逐次的观测值通常是不独立的，而且分析时必须考虑观测资料的时间顺序，这同以前所介绍的观测资料有很大的区别。

时间序列的变化受多种因素的影响，一般可将这些因素分为以下四种：

### 1. 长期趋势 ( $T$ )

长期趋势反映了某种现象在一个较长时间内的发展方向，可以在一个相当长的时间内表现出一种近似直线的持续向上、持续向下或平稳的趋势。也可表现出某种类似指数趋势或其他曲线趋势。粗略地可将“趋势”定义为“均值的长期间变化”。Granger (1966) 定义“均值趋势”为包含波长超过观测时间序列长度的所有频率分量。长期趋势一旦形成，便会延续很长时间，因此对其进行预测研究具有特别重要的现实意义。

### 2. 季节变动 ( $S$ )

季节变动是某种现象受季节变动影响所形成的一种长度和幅度固定的周期波动。许多时间序列如销售量及温度等都显示出年周期的变化。

### 3. 周期变动 ( $C$ )

周期变动也称循环变动，它是由于某些物理原因或经济原因的影响而显示出有固定周期的变化。例如股票价格的变化等，具有明显的周期变动特征。

周期变动有时具有季节变动的特征，比如，像季节变动一样可以预计它缓慢地上下波动，但这里的“周期”一词是用来描述比季节变动更难以预测、更加缓慢的移动。周期长度和峰值都是不确定的，许多周期的平均长度约为 3 至 4 年，有的达 15 年以上。一些学者长期以来一直致力于研究周期的本质和可测性。

对于短期预测，通常会将周期和趋势放在一起考虑，因为此时不可能从短期序列中获取任何有关周期的有用信息。

### 4. 不规则变动 ( $I$ )

不规则变动因素又称随机变动，它是受各种偶然因素的影响所形成的不规则波动。如石油价格受突发事件的影响上涨等。

当将时间序列分解成长期趋势、季节变动、周期变动和不规则变动四个因素后，可将时间序列  $Y$  看成是这四个因素的函数，即

$$Y_t = f(T_t, S_t, C_t, I_t)$$

常用的时间序列分解的模型有加法模型和乘法模型。

加法模型为:  $Y_t = T_t + S_t + C_t + I_t$

乘法模型为:  $Y_t = T_t \times S_t \times C_t \times I_t$

相对而言,乘法模型比加法模型用得更多。在乘法模型中,时间序列值和长期趋势用绝对值表示,季节变动、周期变动和不规则变动用相对值(百分数)表示。

本章主要介绍时间序列分析研究中的序列图、建立模型(指数平滑、综合移动平均)、应用模型、自相关、季节分解、频谱分析、互相关等时间序列分析方法及程序的使用。

SPSS 中进行时间序列分析由主菜单的 Analyze 下拉菜单中的 Time Series 菜单项导出。用鼠标单击主菜单的 Analyze 菜单项,展开下拉菜单,选择 Time Series,显示小菜单,见图 17-1。其中包括:

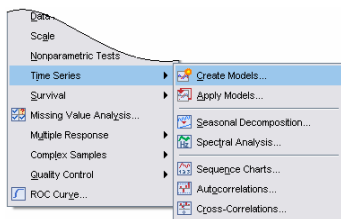


图 17-1 各种时间序列分析过程

- Create models, 建立模型
- Apply models, 应用模型
- Seasonal Decomposition, 季节分解法
- Spectral Analysis, 频谱分析
- Sequence charts, 序列图
- Autocorrelations, 自相关
- Cross-Correlations, 互相关

## 17.1 时间序列的建立和平稳化

在选择上述过程对数据用时间序列模型进行拟合处理前,应先对数据进行必要的预处理和观察。预处理工作分为三个步骤:首先,对有缺失值的数据进行修补,其次将数据资料定义为相应的时间序列,最后对时间序列数据的平稳性进行计算观察。

如果数据文件中存在一个变量,其值是按某一时间间隔采集的,要进行时间序列分析,还需要有一个表明采集时间的日期变量。生成日期变量的方法请见 2.1.3 节的内容。

### 17.1.1 缺失值数据的修补

如果要进行时间序列分析的数据存在缺失值,我们不能采用通常删除的办法来解决,因为这样做的结果将导致原有时间序列周期性的破坏,而无法得到正确的分析结果。

修补缺失值可在 Transform 菜单的 Replace Missing Values 过程中进行。通过单击 Transform→Replace Missing Values 打开缺失值修补对话框,见图 17-2。

1. 从源变量表中选择需修补缺失值的变量,将其送入 New Variables 框。

2. Name 框存储替换缺失值后时间序列的新变量名。

3. Method 下拉列表中提供修补缺失值的方法。共有五种,分别是:

① Series mean, 用整个序列的均数来替换缺失值。为系统默认选项。

② Mean of nearby points, 用相邻若干点的有效值的均数替换缺失值,在 Span of

nearby points 框中输入计算均数所用的相邻点数。

③ Median of near by points, 用若干相邻点的中位数替换缺失值, 在 Span of nearby points 框中设置计算中位数使用的相邻点数。

④ Linear interpolation, 用线性插值法, 即用相邻两点的平均值替换缺失值。如果时间序列的最前或最后数据有缺失值, 则缺失值不被替换。

⑤ Linear trend at points, 用该点的线性趋势替换缺失值。将记录号作为自变量, 时间序列值作为因变量进行回归, 求得该点的预测值。

4. Span of nearby points 框, 设置相应替换方法中需要使用的相邻点数。输入大于等于 2 的整数。如果用时间序列中全部的有效值, 选择 All。

5. Change 按钮, 若替换方法有变化, 单击本按钮可将所作的修改应用于相应的变量。设置完后, 单击 OK 按钮运行。

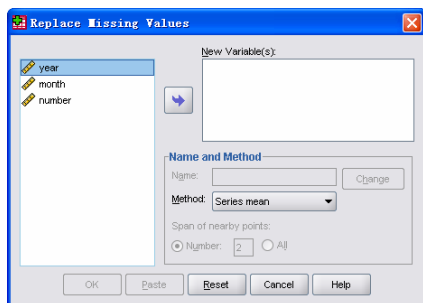


图 17-2 缺失值修补对话框

## 17.1.2 建立时间序列新变量

时间序列分析是建立在序列的平稳的条件上的, 判断序列是否平稳可以看它的均数和方差是否不再随时间的变化而变化, 自相关系数值是否只与时间间隔有关而与所处的时间无关。大多数的时间序列是不平稳的。因此, 首先要识别并将不平稳的时间序列变成平稳的时间序列。

为检验时间序列的平稳性, 经常要用一阶差分、二阶差分, 有时为选择一个合适的时间序列的模型还要对原时间序列数据进行对数转换或平方根转换等。这就需要在已经建立了时间序列的数据文件中, 再建一个新的时间序列的变量。在 SPSS 中 Create Time Series 可根据现有的数值型时间序列变量的函数建立一个新的变量。所建的这些转换值

在许多时间序列的分析程序中经常用到。

判定时间序列的平稳性和趋势特征, 可借助于各种图形来观察, 如序列图、自相关图、频谱图等。

### 1. 建立时间序列新变量的方法

(1) 按 Transform→Create Time Series 顺序展开 Create Time Series 对话框, 见图 17-3。

(2) 选择一个用来建立新变量的数值型变量, 单击向右箭头按钮, 在 New Variable(s) 栏中出现等式, 等号左边是默认的新变量名, 右

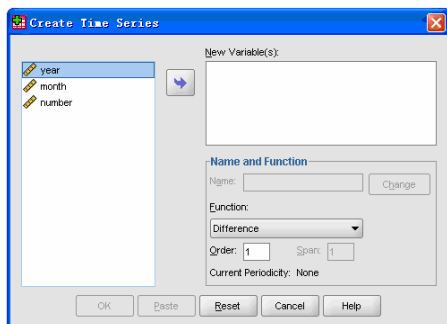


图 17-3 创建时间序列对话框



边是一个在 Function 下拉列表中选定的转换函数, 函数中的参数就是选定的需转换的变量名。

(3) 在 Name and Function 下的 Name 框中, 出现默认的新变量名, 由建立它的变量名前六个字符接下画线和一个有序数字组成。例如变量 sales 对应的新变量名是 sales\_1。如果输入自定义变量名, 单击 Change 按钮确认。新变量保持原变量的值标签。

(4) 在 Function 栏中选择转换函数。提供的有效函数如下:

① Difference 差分函数, 计算时间序列里连续值之间的非季节性差分。Order 是用来计算差分的样品之前的样品数。因为计算一次差分就会丢失一个观测。如果 Order 值为  $n$ , 新时间序列变量中开始  $n$  个值是系统缺失值。例如设置 Order 为 2, 则新变量前两个样品将成为系统缺失值。

② Seasonal difference 季节性差分函数。恒定跨距的序列值之间的差值。跨距取决于当前义的周期。要计算季节差, 必须已经定义了 (data→Define Dates) 包括周期成分的日期变量 (诸如该年的月份)。Order 是用于计算差值的季节周期, 在时间序列开始的系统缺失值的观测数等于周期乘以 Order。例如, 如果当前周期是 12, Order 是 2, 那么新变量前 24 个值将是系统缺失值。

③ Centered moving average 中心移动平均函数。计算以当前值为中心, 在指定跨距范围内, 包括当前值的序列值的平均值。跨距是用于计算平均值的级数。如果跨距是偶数, 移动平均数就是每对非中心均值的均值。如果跨距是  $n$ , 在新变量序列开始和结尾的系统缺失值数就等于  $n/2$ 。例如, 跨度是 5, 在开始和结尾有系统缺失值的观测量数是 2。

④ Prior moving average 前移动平均函数。计算当前值之前的跨距平均值。跨距是用来计算平均值的前面时间序列值的数量, 该序列开始处缺失值的数量等于跨距。

⑤ Running medians 移动中位数函数, 计算以当前值为中心, 在指定跨距范围内 (包括当前值) 的序列值的中位数。计算中位数的时间序列值的数量称为跨距。如果跨距是偶数, 中位数是每对非中心中位数的平均值。对偶数跨距的值和奇数跨距的值而言,  $n$  跨距时间序列起始位置和结束位置含有系统缺失值的样品的数目等于  $n/2$  的整数。例如, 如果跨距是 5, 在开始和结束时的时间序列的系统缺失值的数目是 2。

⑥ Cumulative sum 累积总和函数, 新序列值为到包括当前值的时间序列值的累计和。

⑦ Lag 延迟函数, 当前值取之前第 Order 个样品的值。Order 是当前样品之前的样品的数量。在时间序列的开始处含有系统缺失值的样品数等于 Order 的值。

⑧ Lead 领先函数, 当前值取之后第 Order 个样品的值。Order 是当前样品后的样品数量。在时间序列的结束处含有系统缺失值的样品的数等于 Order 值。

⑨ Smoothing 平滑函数, 用它可以计算原序列的 T4235 平滑序列, 该法又称为 T4253H 平滑法, 最早由 Tukey 提出。对经 T4235 平滑处理后得到的序列, 用 Hanning 权重求移动平均, 从而得到新序列, 故新序列是建立在复合数据平滑器基础上的。它的

功能是经过多步处理将序列中的异常值剔除，使序列平滑。

(5) 单击 OK 按钮，系统运行，可在 OUTPUT 窗口和原数据窗口中，看到运行结果。

## 2. 建立时间序列新变量实例

【例 1】数据 data17-01 为某公司 1973—1999 年的销售额（万元）。用 Lag 函数建立新变量。

(1) 按 Transform→Create Time Series 顺序展开 Create Time Series 对话框，见图 17-3。

(2) 选择 sales 变量，移到 New Variable(s) 栏中。在 Function 选项框中选择函数 Lag。

(3) 在 Name and Function 下的 Name 栏中是默认的新变量名 sales\_1。单击 OK 按钮。

输出如表 17-1 所示。在工作的数据文件中生成 sales\_1 滞后新变量。

表 17-1 运行函数 Lag 时的结果说明

Created Series					
	Series Name	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
		First	Last		
1	sales_1	2	27	26	LAG(sales,1)

在表 17-1 中，第一行从左向右各列依次显示的分别为：新序列变量名、第一个非缺失值观测号、最后一个非缺失值观测号、有效样品数、创建新序列使用的函数。

## 17.2 序 列 图

在构建一个模型以前，为了了解数据的性质，数据是否有季节性波动，可以通过对时间序列绘制连续的样品图来加以判断。

### 17.2.1 序列图过程

按 Analyze→Time Series→Sequence Charts 顺序单击菜单项，展开如图 17-4 所示的 Sequence Charts 对话框。

1. 定义变量：在源变量表中，选定一个或多个满足时间序列要求的或是按有意义顺序排序样品的变量。移到 Variables 框中。

2. 定义时间轴标签变量：在源变量表中，选择一个分类变量，移到 Time Axis Labels 框中。这个变量可以是数值型、字符型或长字符型变量。该变量的值用来标示时间轴。

3. 数据转换：在 Transform 下框中，提供了三种对时间序列或类似的数据进行转换的方法：

(1) Natural log transform，用数据值的自然对数来代替它们本身。这种转换需要所有的值大于 0。

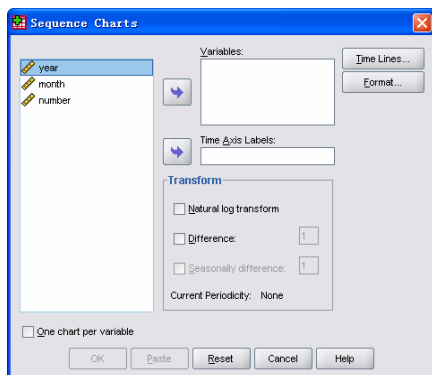


图 17-4 序列图主对话框

(2) **Difference**, 计算两个连续变量之间的差值。输入一个正整数作为差分的阶。一阶差分用当前值减前一个值。但二阶差分是对一阶差分序列做同样的处理, 而不是采用每个值减其之前的两个样品的那个值。

(3) **Seasonally difference**, 通过计算时间跨度相同的两个序列值之间的差值来转换时间序列数据。输入一个正整数作为计算差值的时间周期数。这种转换只有当序列的周期已经定义过时才是有效的 (使用 **Data** 菜单中 **Define Dates** 对话框定义)。

4. 选择图形的输出方式: 选择 **One chart per variable**, 为 **Variables** 框中的每个变量产生一张图。不选择本选项, 则所有的变量绘制在同一张图上。

5. 定义时间轴基准线: 单击 **Time Lines** 按钮, 打开如图 17-5 所示的时间轴参考线对话框, 对绘制的直方图、线图或散点图选择一条任意的刻度线或分类轴线进行定义。选项如下:

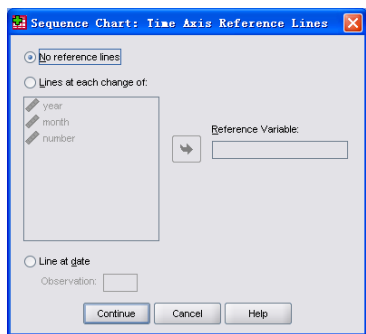


图 17-5 时间轴参考线对话框

(1) **No reference lines**, 输出的图形中没有基准线。

(2) **Lines at each change of**, 基准线会随参考变量的改变而变化。左面的变量列表显示了数据文件中未在主对话框中指定的变量。在变量名列表中选择一个变量并移到 **Reference Variable** 框中作为参考变量。

(3) **Line at date**, 在一个特定的点显示用日期或观察数定义的单个基准线。

① 如果已经定义了日期, 显示用 **Define Dates** 定义的所有日期的标识部分。输入你想要显示基准线处的日期。

② 如果没有定义日期, 则输入想要显示基准线处的参考变量的值。

6. 定义时间轴的格式, 在主对话框中单击 **Format** 按钮, 打开如图 17-6 所示的 **Format** 对话框。选择作图类型以及有关的格式参数。

(1) **Select Time on horizontal axis** 要求水平轴是时间轴, 用垂直轴表示序列值。

(2) **Single Variable Chart(s)** 单一变量图。当只选了一个变量, 或选择了 **One chart per variable** 每个变量一个图, 则对指定的绘图变量可选择做 **Line chart** 线图或 **Area chart**, 面积图。

(3) **Reference line at mean of series**, 在序列均值处画基准线。

(4) **Connect cases between variables**, 一张图中显示多个变量的序列图, 以示变量间和各观察值之间的联系。

单击 **Continue** 按钮, 返回图 17-4 所示主对话框。单击 **OK** 按钮, 运行绘制序列图程序。

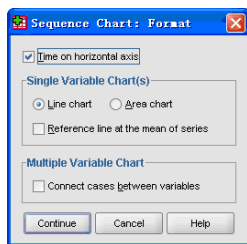


图 17-6 **Format** 对话框

## 17.2.2 序列图应用实例

【例 2】data17-02.sav 是 SPSS16.0 中自带的假设数据文件。包括 1999 年 1 月至 2003 年 12 月 4 年中 85 个地区宽带供货商每月的国家宽带服务用户数量的数据。试用总用户数量序列作序列图。

具体操作步骤如下：

1. 按 Analyze→Time Series→Sequence Charts 顺序，打开如图 17-4 的主对话框。
2. 在源变量表中选 Total Number of Subscribers 作为绘图变量，移到 Variables 框中。
3. 在源变量表中选择 Date 变量作为时间轴变量，移入到 Time Axis Labels 框中。
4. 单击 Time Lines 按钮，在时间轴参考线对话框中，选择 lines at each change of, 在源变量表中，选择供货商 1 的用户数变量，送入 Reference Variable 框中。
5. 选择 line at date, 在 Year 框中输入 2002, 在 Month 框中输入 6。
6. 其他保持系统默认选项，单击 OK 按钮运行，在输出窗中得到如表 17-2、表 17-3 和图 17-7 所示的结果。结果解释如下：

在表 17-2 中，给出了模型的描述，从上到下依次显示的是模型名 MOD\_1、序列或连续数据的数量 1（名称为：用户总数）、转换函数（无）、非季节性差分（0）、季节性差分（0）、季节周期的长度（12）、水平轴的标识（Date）、插入基准线的起始时间（2002 年 6 月）基准线（无）、曲线下的面积（空）。括号中是第二列中对应各行的结果。

在表 17-3 中，给出了样品处理的摘要。从上到下依次显示：序列长度（60）、缺失值数量：用户缺失值（0）、系统缺失值（0）。

表 17-2 模型描述表

Model Description	
Model Name	MOD_2
Series or Sequence	1 用户总数
Transformation	None
Non-Seasonal Differencing	0
Seasonal Differencing	0
Length of Seasonal Period	12
Horizontal Axis Labels	Date_
Intervention Onsets	年=2002, 月=6
Reference Lines	None
Area Below the Curve	Not filled

Applying the model specifications from MOD\_2

表 17-3 样品处理摘要

Case Processing Summary		
	用户总数	
Series or Sequence Length	60	
Number of Missing	User-Missing	0
Values in the Plot	System-Missing	0

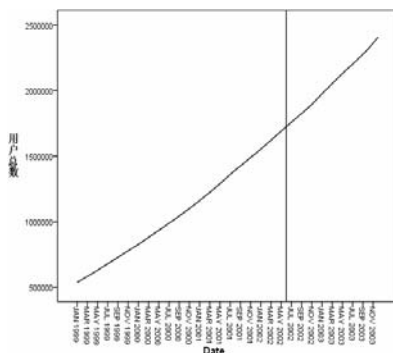


图 17-7 含有基准线的序列图

图 17-7 显示供货商 1 的用户总数的序列图。竖线为基准线，对应的时间为 2002 年 6 月。图中可见：序列展现很平滑的向上趋势，没有一点点季节性波动。总的来说，季节性变化趋势不是数据的显著特征。

当然，如果要对总用户数以外的其他序列做时间序列分析时，在排除有季节性模型的可能性前，应分别地检查各个序列。

## 17.3 建立时间序列模型

时间序列建模程序可对时间序列进行指数平滑、单变量 ARIMA(自回归综合移动平均)估计、多元 ARIMA(或转换函数模型)估计,并产生预报。程序包括为一个或多个因变量时间序列自动识别和估计最适 ARIMA 模型或指数平滑模型的 Expert Modeler,因此,不需要通过反复实验识别一个适当的模型。可以选择自定义 ARIMA 模型或指数平滑模型。

按 Analyze→Time Series→Create models 顺序,展开如图 17-8 所示的建模提示框。

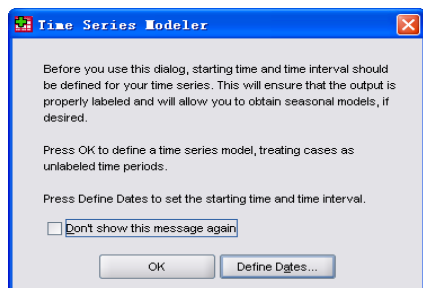


图 17-8 时间序列建模提示框

注意,在使用本对话框之前,应对时间序列定义起始时间和时间间隔,以确保输出标识正确,并且如果需要,可获得季节模型。

下次运行时,若不要显示本信息对话框,可选择 Don't show this messages again。

单击 Define Dates 按钮,可弹出 Define Dates 对话框,在该对话框中,可设置起始时间和时间间隔。具体操作方法参见第 2 章。

如果当前的数据文件已经定义为时间序列时,本对话框不出现。

### 17.3.1 指数平滑与ARIMA模型概述

#### 1. 指数平滑

指数平滑预测方法最先由 C. C. Holt 在 1958 年提出的,它最初只应用于无趋势、非季节作为基本形式的时间序列的分析,后经 Brown、Winter 等统计学家的深入研究和发 展,使指数平滑涉及的数据内部构成更丰富,相应的数据处理方法也更多。指数平滑法的估计是非线性的,其目标是使预测值和实测值间的均方差(MSE)最小。

#### 2. ARIMA 模型

ARIMA 模型广泛应用于时间序列分析的常见模型。它估计非季节和季节平稳性的自回归综合移动平均模型,ARIMA 模型,即著名的 Box-Jenkins 模型。它可以延伸到对包含季节趋势的时间序列进行分析。根据对时间序列特征的预先研究,可以指定三个参数用来分析时间序列,即自回归阶数( $p$ )、差分次数( $d$ )和移动平均阶数( $q$ )。通常模型被写做  $ARIMA(p,d,q)$ 。

Box-Jenkins 方法的第一步是对数据求差分(一阶差分  $\nabla X_t = X_t - X_{t-1}$ , 二阶差分  $\nabla^2 X_t = \nabla X_t - \nabla X_{t-1}$ )直到它是平稳的。这可以通过检查各种差分序列的相关图(包括偏自相关图)直到找出一个“急速”下降于零,并且从此任何季节效应都大大消除的序列来完

成时间序列的随机性、平稳性及季节性的分析。对于非季节数据，通常求一阶差分就足够了。对周期为 12 的季节数据，当季节效应是加性的时，通常可以采用算子  $\nabla_{12}$ ；如果周期效应是乘性的，则可以采用算子  $\nabla_{12}^2$ 。有时算子  $\nabla_{12}$  本身就足够了，不必外加差分。对于季节的数据，可以采用算子  $\nabla_4$  等。

第二步是选定一个特定的模型拟合所分析的时间序列数据。模型识别是 Box-Jenkins 方法中很重要的一环，是否合适的比较标准是：对一般 ARMA 模型体系中的一些特征，分析其理论特征，把这种特定模型的理论特征作为鉴别实际模型的标准，观测实际资料与理论特征的接近程度。最后根据这种分类比较分析的结果来判定实际模型的类型。

该方法的第三步是用时间序列的数据，估计模型的参数，并进行检验，以判定该模型是否恰当。如不恰当，则返回第二步，重新选定模型。

### 17.3.2 选择分析变量

在时间序列建模提示框中，单击 OK 按钮，则关闭提示框，打开如图 17-9 的 Time Series Modeler 对话框 Variables 选项卡。

#### 1. 定义因变量和自变量

在 Variables 框中选定一个或多个变量送到 Dependent Variables 框中，作为因变量。

如果建模需要，在 Variables 框中选定一个或多个变量送到 Independent Variables 框中，作为自变量。

因变量和自变量都应是数值变量。都会被认为是时间序列，也就是说，每一个样品代表了一个时间点，连续的样品通过一个恒定的时间间隔分开。

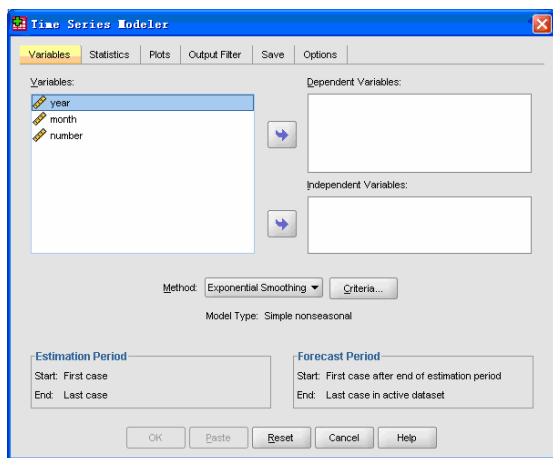


图 17-9 Variables 选项卡

2. 在 Estimation Period 栏显示了估计期的起始和结束位置。估计期即用来估计模型的样品集。默认的是从第一个观测到最后一个观测。如果要改变估计期，使用数据窗口 data 菜单中的 Select cases 功能。在对话框中选择 Based on time 或 case range 设定估计期。改变了的估计期将显示在本栏中。

3. 在 Forecast Period 栏显示了预测期的起始和结束位置。默认的是从估计期结束后的第一个观测开始，到实际数据集的最后一个观测为止。可以在时间序列模型的主对话框 Options 选项卡中改变预测期。改变后本栏显示新设定的预测期。

例如，时间序列是从 1999 年 1 月开始到 2003 年 12 月结束的 4 年数据。这也是默认估计期的范围。如果定义估计期从 2000 年 1 月开始到 2002 年 12 月止。则默认的预测期从 2003 年 1 月开始，到 2003 年 12 月止。

#### 4. 确定建模方法

在 Method 的下拉列表中, 共有 3 个选择项: Expert Modeler (专家建模) 是系统默认的建模方法、Exponential Smoothing (指数平滑法)、ARIMA (自回归综合移动平均法)。

##### 17.3.2.1 Expert Modeler 专家建模

Expert Modeler 会自动地为每个因变量序列找到最佳拟合模型。

在主对话框 Method 下拉列表中选择 Expert Modeler 专家建模, 单击 Criteria 按钮, 打开如图 17-10 所示的 Expert Modeler Criteria 对话框。

#### 1. Model 选项卡

(1) 在 Model Type 栏选择模型的类型, 有 3 个有效选择项。

① All models, 系统默认选项。同时考虑 Exponential Smoothing 指数平滑法和 ARIMA 自回归综合移动平均法, 程序会自动识别用哪个作为拟合时间序列的最佳模型。

② Exponential smoothing models only, 只对时间序列采用指数平滑法进行估计。

③ ARIMA models only, 只对时间序列采用自回归综合移动平均法进行估计。

④ Expert Modeler considers seasonal models, 专家建模考虑季节模型。只有当前数据文件已定义周期时才有效。选择本项, Expert Modeler 同时考虑季节性和非季节性模式。如果未选择此选项, Expert Modeler 只考虑非季节性模式。

在 Current Periodicity 后面, 显示已定义的周期。如果没有定义周期, 则显示值 None。

(2) Events 栏。选择任何自变量都被当作事件变量。事件变量值为 1 的样品表明该时期的因变量序列被期望受事件影响。1 以外的值表明没有影响。

2. Outliers 选项卡如图 17-11 所示, 可以选择自动检测异常值的类型。

(1) Detect outliers automatically, 自动检测异常值。默认不自动检测异常值。



图 17-10 专家建模标准模型选项卡

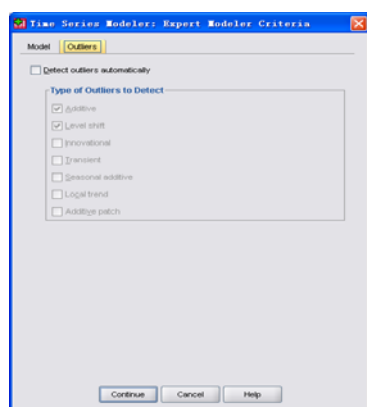


图 17-11 专家建模判断异常值选项卡

(2) Type of Outliers Detect 栏选择一项或多项异常值类型:

① Additive, 影响单个观察值的异常值。例如, 一个数据编码错误可能被认为是一



个加性异常值。

② Level shift, 在一个特定的序列点（异常值出现的时间点）开始, 所有观察值用其自身值加上一个常量来转换, 该常量等于特定的序列点处的异常值与其真值的偏差。具体可参阅 SPSS 中的 LS (TSMODEL algorithms) 算法。

③ Innovational, 在一个特定的序列点, 从异常值起增加噪声项的作用。对平稳序列, 创新异常值 (Innovational Outlier) 只影响少数观察值。但对非平稳序列, 创新异常值可能会影响在一个特定的序列点开始的每个观察值。

④ Transient, 一个影响衰减指数直到 0 的异常值。

⑤ Seasonal additive, 周期性加性异常值。影响一个特定的观察值和其后由一个或多个周期隔开的所有的观察值。所有这些观察值受到的影响相同。如果从某年开始的各个 1 月份销售额都较高, 则周期性加性异常值可能会发生。

⑥ Local trend, 局部趋势异常值。在一个特定的序列点开启局部趋势的异常值。

⑦ Additive patch, 两个或两个以上连续的加性异常值组。选择本异常值类型会导致对这些序列点以外的单个加性异常值的检测。

单击 Continue 按钮, 返回到图 17-10 所示的主对话框。

### 17.3.2.2 Exponential Smoothing 指数平滑方法

只有仅指定了因变量时, 在 Method 下拉列表中才有指数平滑方法。

主对话框中 Method 下拉列表中选定本项, 单击 Criteria 按钮, 打开如图 17-12 所示的 Exponential Smoothing Criteria 对话框。

1. Model Type 栏中选择模型的类型: Nonseasonal 非季节模型、Seasonal 季节模型。

(1) 在 Nonseasonal 非季节模型中有 4 个选项:

① Simple, 简单模型适用于无趋势或无季节因素影响的时间序列。唯一的平滑参数是水平。简单指数平滑同具有零阶自回归、一阶差分、一阶移动平均的 ARIMA 模型极其相似, 并且没有常量。

② Holt's linear trend, Holt 线性趋势模型适用于有线性趋势和无季节性因素影响的时间序列。其平滑参数为水平和趋势, 它不受彼此值的约束。Holt 模型比 Brown 模型更普通, 而且计算一个较大的时间序列时要花更长的时间。Holt 指数平滑模型同具有零阶自回归、二阶差分、二阶移动平均的 ARIMA 模型极其相似。

③ Brown's linear trend, Brown 线性趋势模型适用于有线性趋势和无季节因素影响的时间序列。其平滑参数为水平和趋势, 且假定它们

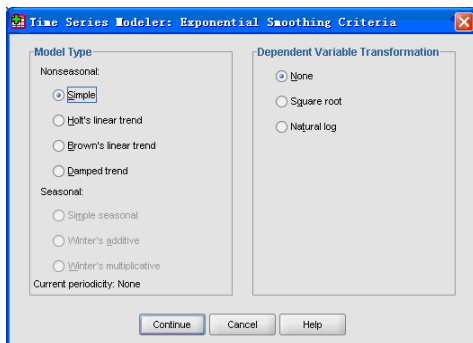


图 17-12 指数平滑标准模型选项卡



相等。Brown 模型因此是 Holt 模型的特例。Brown 指数平滑模型同具有零阶自回归、二阶差分、二阶移动平均的 ARIMA 模型很相似,同时第二阶移动平均系数等于第一阶系数一半的平方。

④ Damped trend, 衰减趋势模型适用于线性趋势正在消失且无季节因素的时间序列。其平滑参数为水平、趋势和阻尼趋势(damping trend)。阻尼指数平滑与具有一阶自回归、一阶差分、二阶移动平均的 ARIMA 模型很相似。

(2) 在 Seasonal 季节性模型项中,有 3 个选项:

① Simple seasonal, 简单季节性模型适用于无趋势和季节因素影响在时间上是常量的时间序列。其平滑参数是水平和季节。简单季节性指数平滑最类似于具有零阶自回归、一阶差分、一阶季节差分、一阶、 $p$  阶及  $p+1$  阶移动平均的 ARIMA 模型,其中  $p$  是在一个季节间隔中的周期数(如每月数据,  $p=12$ )。

② Winters' additive 温特加性模型适用于有线性趋势且不依赖于序列水平季节性影响的时间序列。其平滑参数是水平、趋势和周期。Winters 加性指数平滑与具有零阶自回归、一阶差分、一阶季节差分、以及  $p+1$  阶移动平均的 ARIMA 模型极其相似。其中  $p$  是在一个季节间隔中的周期数(如每月数据,  $p=12$ )。

③ Winters' multiplicative 温特积性模型适用于有线性趋势且依赖于序列水平的周期性影响的时间序列。其平滑参数是水平、趋势和周期。Winters 积性指数平滑同任何 ARIMA 模型都不相似。

在 Current Periodicity 后面的整数是当前的数据文件定义的周期。例如,年周期为 12,每个样品表示一个月。如果没有设定周期,则显示的值为 None。季节模型需要设置周期。可以在 Define Dates 对话框中设置。

2. 在 Dependent Variable Transformation 中定义对时间序列的转换方法。有 3 个选项供选择,在建模之前,可以对每个因变量实施转换。

① None, 对时间序列不实施转换。

② Square root, 对时间序列用平方根转换。

③ Natural log, 对时间序列用自然对数转换。

单击 Continue 按钮,返回到图 17-10 所示的对话框。

### 17.3.2.3 ARIMA 自回归综合移动平均模型。

主对话框的 Method 下拉列表中选择 ARIMA,单击 Criteria 按钮打开如图 17-13 所示的 ARIMA Criteria Model 选项卡。

1. 在 Model 选项卡中指定自定义模型结构。

(1) ARIMA Orders 栏

① Structure 结构表的单元格中需要输入非负整数,以便定义 ARIMA 模型的构成。对自回归和移动平均构成,值代表最大的阶数。所有比最大阶数小的阶数都包含在模型中。例如,如果指定 2,则模型包括 2 阶和 1 阶。Seasonal 栏中只在当前数据文件已定

义周期时有效。

- Autoregressive (p), 设置用序列的先前值来预测现值的自回归的阶数。自回归的阶数 2 指定用序列过去 2 个时间周期的值来预测现值。

- Difference (d), 指定用于时间序列的差分转换的阶数。当存在趋势时 (含有趋势的时间序列典型地是非平稳的, 假设 ARIMA 模型是平稳的), 差分是必要的, 差分可用来消除趋势的影响。差分的阶数同时间序列趋势的程度相对应, 一阶差分说明线性趋势, 二阶差分说明二次趋势, 等等。

- Moving Average (q), 指定模型中移动平均的阶数。移动平均阶数定义了用原先值同序列平均值的偏差来预测当前值。例如, 1 阶和 2 阶移动平均说明当预测时间序列的当前值时, 要考虑最后两个时间周期中的每一个值同时间序列的平均值的偏差。

② Seasonal Orders, 季节性自回归、移动平均以及差分构成同它们在非季节序列对应项含义相同。对于季节性阶数, 时间序列当前值受由一个或几个季节周期隔开的先前序列值影响。例如, 对于每月数据 (季节性周期为 12), 季节阶数 1 意味着当前的序列值受先于当前值 12 个周期的序列值影响。季节阶数 1, 对每月数据, 等于指定了 12 的非季节阶数。

Current Periodicity, 显示当前数据文件定义的周期, 是一个整数。例如, 12 代表每年的周期, 每个样品表示一个月。如果没有设置周期, 则显示值 None。季节模型需要周期。读者可以在 Define Dates 对话框中设置周期。

(2) Dependent Variable Transformation 栏指定各因变量进入模型之前的转换

- ① None, 对时间序列不进行任何转换。
- ② Square root, 对时间序列进行平方根转换。
- ③ Natural log, 对时间序列进行自然对数转换。

(3) Include constant in model, 除非你确信全部序列值的平均数为 0, 否则在模型中应该包含常数项。即应该选择此项。当使用差分时, 建议在模型中不要常数项。

2. 单击 Outliers 选项卡, 在如图 17-14 所示的对话框中选择对异常值的处理方法。

- (1) Do not detect outliers or model them 栏, 系统默认异常值不被侦查也不进入模型。
- (2) Detect outliers automatically 栏, 自动对异常值进行侦查的选项。

在此可以选择要侦查的一个或多个异常值的类型。提供选择的异常值类型有 Additive、Level shift、Innovational、Transient、Seasonal additive、Local trend、Additive patch,

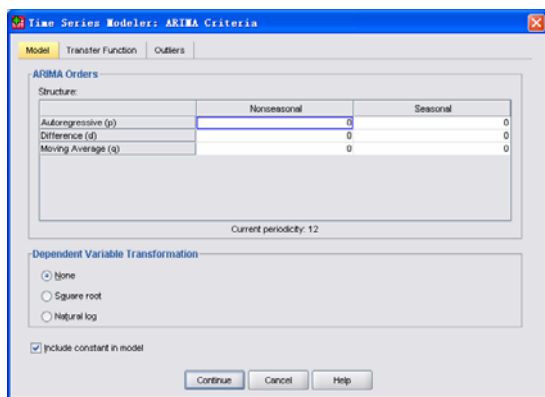


图 17-13 ARIMA Criteria Model 选项卡

有关这些选项的说明，参见 17.3.2.1 节中的相关内容。

(3) Model specific time points as outliers 栏，设置特定的时间点作为异常值。每个异常值占 Outlier Definition 表中的一行。在这行的各单元格中输入时间点的数值。

① Type 列的下拉列表中选择异常值的类型。支持的类型有 Additive（系统默认项）、level shift、innovational、transient、seasonal additive 以及 local trend。

② 在 type 前的两列表头根据定义的周期显示不同的内容。例如 Year、Month 等。在各行相应位置输入表达异常值时间点的数值。注意，如果没有定义日期变量，则 Outlier Definition 表显示单列。为说明一个异常值，输入异常值样品所处的行数。见 Data Editor 窗口。如果表中出现 Cycle 列，它与当前数据文件中 CYCLE\_变量的值有关。

### 3. 在 Transfer Function 选项卡中定义自变量的转换函数

只有在主对话框 Variable 选项卡中指定了因变量和自变量，并在 Methd 下拉列表中指定了 ARMA 方法，单击 Criteria 按钮打开的对话框中才会有 Transfer Function 选项卡。

单击 Transfer Function 选项卡标签，打开如图 17-15 所示的 Outliers 对话框。在这个选项卡中，对在 Variables 选项卡中指定的一个或多个自变量指定转换函数。转换函数是指定用自变量（预测变量）的过去值来预报因变量的将来值的方法。

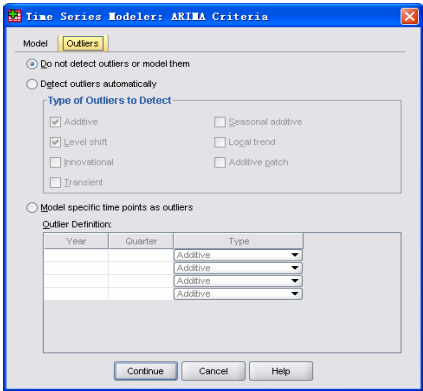


图 17-14 侦查异常值的选项卡

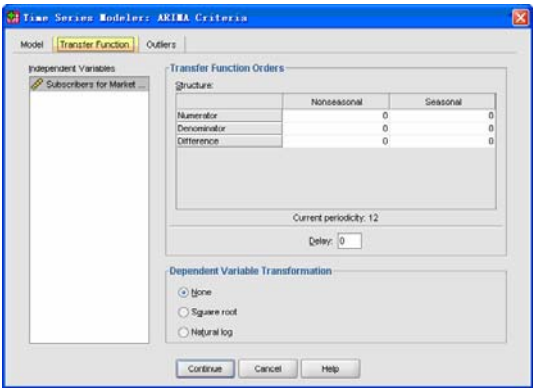


图 17-15 自变量转换选项卡

### (1) Transfer Function Orders 栏

在 Structure 下表格的单元格中输入转换函数不同成分的值。所有值必须是非负的整数。对 Numerator（分子）和 Denominator（分母），输入的值代表了最大的阶数。所有比指定值小且大于等于 0 的阶数都会包括在模型中。例如，如果在 Numerator 后面的单元格中输入 2，那么模型包括 2、1 和 0 阶。如果在 Denominator 后面的单元格中输入 3，那么模型包括 3、2 和 1 阶。如果 Seasonal 列没有被激活，那是因为当前工作的数据文件中没有定义周期。

① Numerator，指定转换函数分子的阶数。用选择的自变量（预测变量）序列中指

定阶数的先前值去预测因变量的当前值。例如，1 阶的分子指定用过去一个时间周期的自变量序列的值，以及自变量序列的现值，预测各个因变量序列的现值。

② Denominator，指定转换函数分母的阶数。指定选择的自变量（预测变量）序列的指定阶数的先前值与时间序列均数的偏差来预测因变量的当前值。例如 1 阶的分母指定在预测各因变量序列的现值时，要考虑自变量序列过去一个时间周期的平均值的偏差。

③ Difference，在估计模型前，指定用于选择的自变量（预测变量）序列差分的阶数。当存在趋势及要消除它的影响时，需要用差分。

④ Seasonal Orders，季节性的分子、分母和差分成分同非季节性的分子、分母和差分成分发挥同样的作用。对季节性阶数，当前序列值受到由一个或多个季节周期分开的先前序列值的影响。例如，对每月的数据（季节周期 12）而言，季节阶数 1 意味着当前序列值受到先于当前值 12 个周期的序列值的影响。

### (2) Current Periodicity 栏

指出在工作的数据集中定义过的当前周期（如果有的话）。

### (3) Delay 栏

通过指定一个间隔数设置延迟，促使自变量的影响相应延后。例如，如果设置延迟为 5，则在时间  $t$  的自变量的值不影响预报，而  $t+5$  过后的自变量的值对预报有影响。

### (4) Dependent variable transformation 栏选择对自变量作转换的方法

- None，不做转换。它是默认选项。
- Square root，用平方根转换。
- Natural log，用自然对数转换。

## 17.3.3 选择统计量

在主对话框中单击 Statistics 选项卡，得到图 17-16 时间序列模型统计量对话框。选择输出建模结果的选择项。

1. Display fit measures, Ljung-Box statistic, and number of outliers by model，显示的表格中包括选择的拟合程度的测度、Ljung-Box 值以及各模型的异常值数。

2. Fit Measures 栏中选择拟合测度。

(1) Stationary R-squared 平稳  $R^2$ ，将模型的平稳部分和简单平均模型进行比较。当有趋势或季节模式时，本测度比普通  $R^2$  更好。平稳  $R^2$  方值范围是负无穷大到 1。负值意味着在考虑中的模型比基准模型更糟。正值意味着在考虑中的模型比基准模型更好。

(2) R-squared 即  $R^2$ ，由模型解释的时间序列中的总变异比例的估计。当时间序列是平稳序列时，本测度极有用。 $R^2$  值的范围是负无穷大到 1。负值表示考虑中的模型不如基准模型。正值表示优于基准模型。

(3) RMSE 均方根误差。因变量序列同其模型预测值间差异程度的测度，用与因变量序列相同的单位来表示。

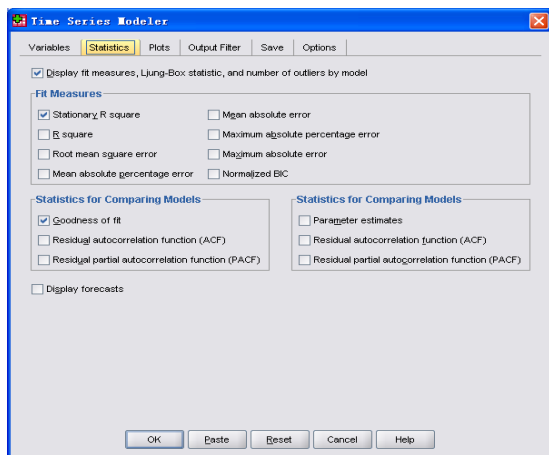


图 17-16 时间序列模型 Statistics 选项卡

变量来表示。像 MaxAPE 一样，对预测设想一个最坏的结果时，本测度很有用。最大绝对误差和最大绝对百分比误差可以发生在不同的序列点。例如，当一个大的序列值的绝对误差比一个小序列值的绝对误差稍大时，最大绝对误差将发生在较大的序列值处，最大绝对百分比误差将发生在较小的序列值处。

(8) Normalized BIC，标准化贝叶斯信息准则。它是试图说明模型复杂性的对模型整体拟合的综合测度，是建立在均方误差基础上的得分，并包括在模型中参数的数量和序列在长度上的损失。这个损失降低了有更多参数的模型的优势，从而使统计量更容易对同一序列的不同模型进行比较。

3. Statistics for Comparing Models 栏中选择比较模型的统计量。本组选项控制包括所有估计模型的统计计算的表格的显示。每个选项会产生一个单独的表。

(1) Goodness of fit，拟合优度统计量，产生平稳  $R^2$ 、 $R^2$ 、均方误差的平方根、均数绝对百分比误差、均数绝对误差、最大绝对百分比误差以及标准化贝叶斯信息准则的摘要统计表和百分比表。

(2) Residual autocorrelation function (ACF)，产生所有估计模型残差自相关的摘要统计表和百分比表。

(3) Residual partial autocorrelation function (PACF) 产生所有估计模型残差偏自相关的摘要统计表和百分比表。

4. Statistics for Individual Models 栏，选择包括每个估计模型详情的表格。每个选项产生一个单独的表格：

(1) Parameter estimates 参数估计，为每个估计模型显示一张参数估计表。为指数平滑模型和 ARIMA 模型显示单独的表。如果异常值存在，也会显示在一个单独的表中。

(4) MAPE 平均绝对百分比误差。因变量序列同它的模型预测值间差异程度的测度。由于它不依赖于使用的单位，因此它能用来比较不同单位的序列。

(5) MAE，平均绝对误差。序列同它的预测模型水平间差异程度的测度。MAE 采用原先序列的单位。

(6) MaxAPE，最大绝对百分比误差。用百分比表示的最大预测误差。对于预测设想一个最坏的结果时，本测度很有用。

(7) MaxAE，最大绝对误差。最大预测误差，用因变量序列的相同单位来表示。

(2) Residual autocorrelation function (ACF) 为每个估计模型显示一张滞后的残差自相关表, 包括自相关的置信区间。

(3) Residual partial autocorrelation function (PACF) 为每个估计模型显示一张滞后的残差偏自相关表, 包括偏自相关的置信区间。

5. Display forecasts 选项为每个估计模型显示一张模型预测和置信区间表。预测期可在 Options 选项卡中设置。

### 17.3.4 Plots图形

单击 Plots 选项卡, 打开如图 17-17 所示的 Time Series Modeler Plots 对话框。在本选项卡中, 选择建模结果图。

1. Plots for Comparing Models 栏中选择表现模型拟合程度的图形, 其中前 8 个选择项对应 Statistics 选项卡中 Fit Measure 中的 8 个统计量。每个选择项单独产生一个图形。可以多项同时选择。供选择的图形有 Stationary R-square、R-square、Root mean square error、Mean absolute percentage error、Mean absolute error、Maximum absolute percentage error、Maximum absolute error、Normalized BIC。统计量解释见 17.3.3 节。此外还有两个选择项:

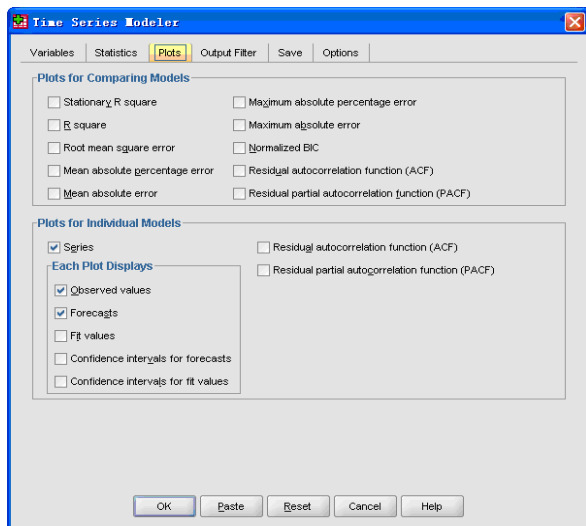


图 17-17 Time Series Modeler Plots 选项卡

- Residual autocorrelation function (ACF), 在图形中包含残差自相关函数。
- Residual partial autocorrelation function (PACF), 在图形中包含残差偏自相关函数。

2. Plots for Individual Models 栏中的选项是针对单个模型的。

(1) Series 对每个估计模型产生预测值图。每个图包含的内容可以选择下列选项:

- ① Observed values, 在图形中包含因变量序列的观察值。
- ② Forecasts 图中显示预测期中的模型预测值。
- ③ Fit values, 在图形中包含估计期的模型预测值。
- ④ Confidence intervals for forecasts, 在图形中包含预测期的置信区间。
- ⑤ Confidence intervals for fit values, 在图形中包含估计期的置信区间。

3. Residual autocorrelation function (ACF), 为各估计模型显示残差自相关图。

4. Residual partial autocorrelation function (PACF), 显示各估计模型残差偏自相关图。

### 17.3.5 输出项目的过滤

单击 Output Filter 选项卡, 打开如图 17-18 所示的 Output Filter 对话框。对话框中是对估计模型子集的表和图的输出进行限制的选项。

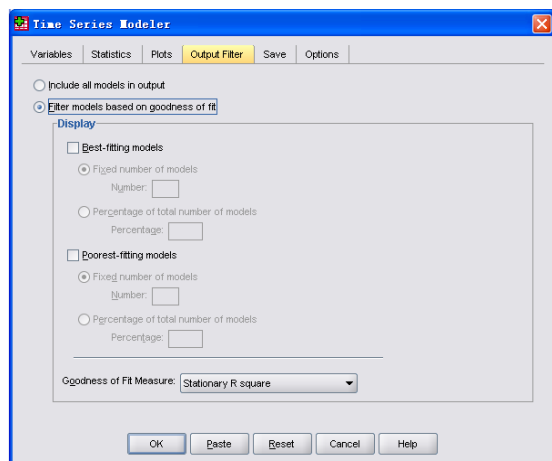


图 17-18 Output Filter 选项卡

1. Include all models in output, 系统默认输出包含所有的估计模型。

2. Filter models based on goodness of fit, 根据拟合优度限制模型的输出。

(1) Best-fitting models, 输出最佳拟合模型, 选择此项还要输入限制参数:

① Fixed number of models, 指定显示  $n$  个最佳拟合模型。如果  $n$  大于估计模型的数量, 则显示所有的模型。

② Percentage of total number of models, 显示模型中拟合优度最高的前  $n\%$  个模型。

(2) Poorest-fitting models, 输出最差拟合模型, 选择此项还要输入限制参数:

① Fixed number of models, 指定显示  $n$  个最差拟合模型的结果。如果指定的数量超过估计模型的数量, 则显示所有的模型。

② Percentage of total number of models, 显示模型中拟合优度最低的最后  $n\%$  个模型。

(3) Goodness of Fit Measure, 拟合优度的测度

选择以上两种过滤方式, 还需要指定用于过滤模型的拟合优度测度。在下拉列表中, 共有 8 个选项, 系统默认为平稳  $R^2$ 。它们分别是 Stationary R-square、R-square、Root mean square error、Mean absolute percentage error、Mean absolute error、Maximum absolute percentage error、Maximum absolute error、Normalized BIC。有关这些选项的说明, 参见 17.3.3 节中的相关内容。

### 17.3.6 保存新变量

单击 Save 选项卡, 进入如图 17-19 所示的 Time Series Modeler Save 对话框。在本对话框中, 指定要保存在当前工作的数据文件的新变量和保存到外部文件的选择项。

#### 1. 保存新变量

在 Save Variables 栏中, 可以在当前工作的数据文件中存储模型预测值、置信区间以及残差的新变量。每个因变量序列建立与自身有关的新变量, 每个新变量包含估计和预测期的值。如果预测期超过因变量序列的长度, 则增加新样品。通过为以下的各个统计



量选择相关的“Save”选项，选择存储新变量。在系统默认情况下，不保存新变量。

① Predicted Values，存储模型预测值。

② Lower Confidence Limits，存储预测值置信区间的下限。

③ Upper Confidence Limits，存储预测值置信区间的上限。

④ Noise Residuals，存储模型残差。如果对因变量进行了转换，例如用自然对数进行转换，存储的是转换序列的残差。

⑤ Variable Name Prefix，指定新变量名的前缀，或留下默认的前缀。变量名由前缀、与因变量有关的名字以及模型标识符组成。如果必须避开变量名的冲突，则变量名用扩展名。前缀必须符合有效变量名的命名规则。

## 2. 对输出模型文件进行命名

对所有估计模型的模型说明被输出到 XML 格式的指定文件中。存储的模型可以对更多的数据，使用 Apply Time Series Models 程序，获取新的预测。

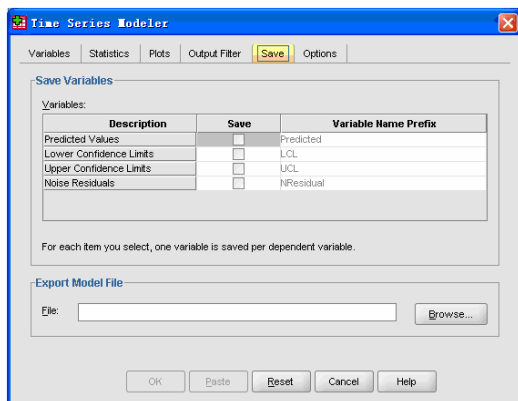


图 17-19 Time Series Modeler Save 选项卡

## 17.3.7 建模的其他选择项

单击 Option 选项卡，进入如图 17-20 所示的 Time Series Modeler Options 对话框。在

选项卡中，可以设置预测期、指定缺失值的处理、设置置信区间的宽度、为模型的标识指定一个自定义的前缀，以及为自相关设置滞后显示的数量。

### 1. 定义预测期

在 Forecast Period 栏中选择预测期的位置。通常预测期在估计期（用来确定模型的样品集）结束后的第一个样品开始一直到当前工作数据文件中的最后一个样品或使用用户指定的日期时结束。在默认状态下，估计期结束在当前工作数据文件中的最后一个样品，但它可以在 Select Cases 对话框中通过选择 Based on time or case range 选项来改变。

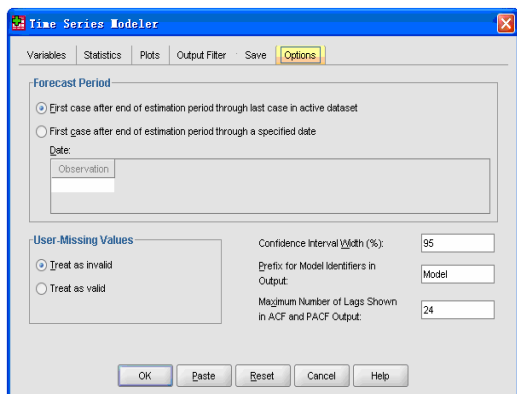


图 17-20 时间序列模型 Options 选项卡



① First case after end of estimation period through last case in active dataset, 估计期结束后的第一个观测到实际数据集的最后一个观测作为预测期。当估计期先于当前工作数据文件的最后一个样品前结束, 又要预测到最后一个样品时选择本项, 本选项主要用来为延续期产生预测, 允许同期的模型预测值同实际值的子集进行比较。

② First case after end of estimation period through a specified date, 估计期结束后的第一个观测到指定的日期。本选项要求指出预测期的结束点。本选项主要用来产生超出实际序列最后范围的预测。在 Date 栏所有单元格输入日期值。如果当前工作的数据文件中没有定义日期说明, 那么 Date 栏中显示 Observation。输入预测期结束点相应样品的行数(同在 Data Editor 中显示的一样)。

在 Date 栏若显示 Cycle 与当前工作的数据文件里 CYCLE\_变量的值有关。

2. 在 User-Missing Values 的选项中选择对用户缺失值的处理方法。其选项有:

① Treat as invalid, 用户缺失值当作系统缺失值处理。

② Treat as valid, 用户缺失值当作有效值处理。

3. 用户缺失值的处理

在 User Missing Values 栏可以指定对缺失值 Treat as invalid 作为非法值处理, 也可以指定 Treat as valid 做合法值处理。做合法值处理要注意以下三点:

① 估计期内因变量缺失值样品包含在模型中。缺失值的特殊处理取决于估计模型。

② 自变量在估计期内有缺失值, 则发出警告。对专家建模法, 可以估计没有自变量, 或自变量包含缺失值的模型。自定义 ARIMA 模型要求自变量不能包含缺失值。

③ 如果在预测期中任一个自变量有缺失值, 程序发出警告且据程序所能进行预测。

4. 定义置信区间

在 Confidence Interval Width (%) 后框中, 可以指定任意小于 100 的正数。为模型预测计算置信区间和残差自相关。在默认状态下, 使用 95% 的置信区间。

5. 为输出中的模型标识定义前缀

在 Prefix for Model Identifiers in Output 后框中, 可以输入前缀或保留模型的默认名。在 Variables 对话框中指定每个因变量的估计模型。模型用唯一名字区分, 名字由定制的前缀和整数后缀组成。

6. 定义 ACF 和 PACF 输出中显示的最大滞后数

在 Maximum Number of Lags Shown in ACF and PACF Output 后框中, 可以设置在表和自相关与偏自相关图中显示的最大滞后数。

### 17.3.8 时间序列分析实例

【例 3】仍以 17.2.2 节中存放在 data17-02.sav 中的 1999—2003 年 85 个地区宽带供货商每月的国家宽带服务用户数量的数据文件为例, 试用 Expert Modeler 对每个地区宽带供货商每月的国家宽带服务用户数量的数据进行时间序列分析。

注意：变量标签名须是英文，否则不能做超出数据文件长度的预测。

打开 data17-02.sav 文件。具体操作步骤如下：

1. 按 Analyze→Time Series→Create models 顺序，打开 Time Series Modeler 对话框 Variables 选项卡，见图 17-9。

2. 在源变量表中选择 Market\_1~Market\_85，即供货商 1 的用户数到供货商 85 的用户数共 85 个变量，并将其移入到 Dependent Variables 框中。要求拟合 85 个模型。

3. 在 Method 的下拉列表中选择 Expert Modeler 专家建模。

4. 单击 Criteria 按钮，打开 Expert Modeler Criteria 对话框，见图 17-10。虽然当前周期是 12，做出序列图，我们可以知道 85 个供货商的用户数的时间序列中同样不存在季节性因素的影响，所以可以不考虑季节模型，故在 Model Type 选项组中，撤销 Expert Modeler considers seasonal models 选项。这样使用 Expert Modeler 研究时可以减少建模占用的计算机存储空间和计算时间。

5. 单击 Continue 按钮，返回 Time Series Modeler 对话框 Variables 选项卡。

6. 单击 Options 选项卡，进入 Time Series Modeler Option 对话框，见图 17-20。

在 Forecast Period 下的选项中，选择 First case after end of estimation period through a specified date。在 Date 项下的 year 下面输入 2004，在 month 下面输入 3。设置的预测期将从 2004 年 1 月到 2004 年 3 月。其他采用系统默认选项。

7. 单击 Statistics 选项卡，进入 Time Series Modeler Statistics 对话框，见图 17-16。

选择 Display forecasts，为每个因变量序列产生一张预测值表。

在 Statistics for Comparing Models group 栏中，Goodness of fit 默认选项产生拟合统计量汇总表，如用所有模型计算的  $R^2$ 、均数绝对百分比误差以及标准 BIC。

8. 单击 Plots 选项卡，进入 Time Series Modeler Plots 对话框，见图 17-17。

由于我们对存储预测值为一个新变量比产生预测图更感兴趣，所以，在 Plots for Individual Models 选项中，撤销 Series 选项。禁止为每个模型产生序列图。

在 Plots for Comparing Models 栏中选择 Mean absolute percentage error 均数绝对百分比误差(MAPE)和 Maximum absolute percentage error 最大绝对百分比误差(MaxAPE)。

绝对百分比误差是因变量序列同它的模型预测水平间有多少差异的测度。通过检查所有模型的均数和最大值，可以得到在预测中有不确定性的迹象。于是查看百分比误差的概要图，而不是绝对误差图是明智的，因为因变量序列代表大小不同的市场的用户数。

9. 单击 Save 选项卡进入 Time Series Modeler Save 对话框，见图 17-19。在 Save Variables 栏 Variables 表中选择保存 Predicted Values，并使用 Predicted 作为变量名前缀。

单击 Browse 按钮，设置存储位置，输入 XML 格式文件名 model17-02。

10. 单击 OK 按钮，提交运行。得到输出结果见表 17-4~表 17-7 和图 17-21~图 17-23。

11. 结果解释

表 17-4 模型描述显示了最佳拟合各供货商用户数的时间序列模型。

第一列显示的是供货商用户数的模型编号，第二列显示其对应的最佳拟合模型名称，参见 17.3.2.2、17.3.2.3 节的有关内容。

图 7-21 中给出的直方图显示了所有模型的平均绝对百分比误差的频数条形图。由于大部分模型的平均绝对百分比误差在 0.8~1.0 之间，它表明所有模型显示了大概 1% 的平均不确定性。

图 17-22 给出的直方图显示了所有模型的最大绝对百分比误差，它对设想预测的最坏情况方案是有用的。它显示每个模型的最大百分比误差落在 1%~5% 的范围中。

这些值能否代表可接受的不确定性的量呢？这要取决于个人的商业直觉，因为可接受的风险将因问题而改变。

表 17-5 列出了模型拟合的各种统计量，从左到右各列的表头依次为拟合的统计量、平均数、标准误差、最小值、最大值、百分比（5%、10%、25%、50%、75%、90%、95%），第一列列出 8 种拟合优度测度。其余各列是这些拟合优度测度统计量的计算结果。

表 17-4 模型描述

Model Description			Model Type
Model ID	Subscribers for Market	Model	
Model_1	Subscribers for Market 1	Model_1	Brown
Model_2	Subscribers for Market 2	Model_2	ARIMA(1,1,0)
Model_3	Subscribers for Market 3	Model_3	Brown
Model_4	Subscribers for Market 4	Model_4	ARIMA(0,1,3)
Model_5	Subscribers for Market 5	Model_5	Hot
Model_6	Subscribers for Market 6	Model_6	ARIMA(1,1,0)
Model_7	Subscribers for Market 7	Model_7	Brown
Model_8	Subscribers for Market 8	Model_8	ARIMA(0,1,0)
Model_9	Subscribers for Market 9	Model_9	Brown
Model_10	Subscribers for Market 10	Model_10	ARIMA(1,1,0)
Model_11	Subscribers for Market 11	Model_11	Brown
Model_12	Subscribers for Market 12	Model_12	Brown
Model_13	Subscribers for Market 13	Model_13	ARIMA(1,2,0)

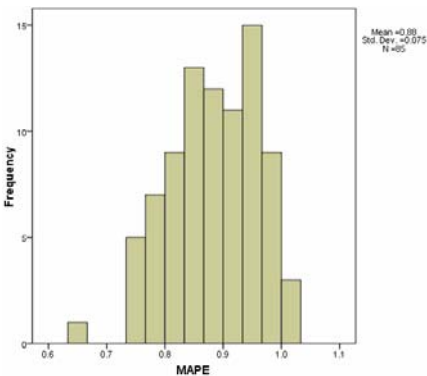


图 17-21 均数绝对百分比误差频数图

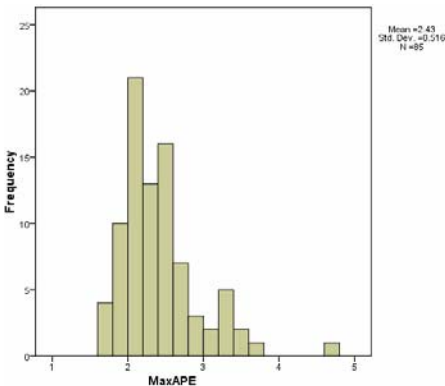


图 17-22 最大绝对百分比误差频数图

表 17-5 模型拟合

Model Fit											
Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.183	.144	-2.52E-15	.629	-6.97E-16	6.71E-17	.068	.188	.247	.376	.478
R-squared	.999	.000	.998	1.000	.998	.999	.999	.999	.999	1.000	1.000
RMSE	177.951	138.821	42.088	737.540	51.471	58.625	90.316	135.016	195.507	402.111	480.796
MAPE	.883	.075	.647	1.007	.748	.775	.831	.885	.946	.980	.991
MaxAPE	2.426	.516	1.708	4.765	1.798	1.937	2.097	2.307	2.604	3.260	3.456
MAE	139.634	106.313	34.033	589.708	40.608	46.836	71.445	107.377	150.785	316.144	343.171
MaxAE	456.170	378.664	108.378	1813.295	117.841	137.366	206.780	340.308	517.760	1.070E3	1.356E3
Normalized BIC	10.003	1.316	7.618	13.345	7.994	8.256	9.111	9.902	10.619	12.062	12.445

通常重点关注两个统计量：MAPE 平均绝对百分比误差和 MaxAPE 最大绝对百分比误差。例如，模型 95% 的 MaxAPE 的值等于 3.456%。所有模型的 MAPE 从最小值 0.647% 到最大值 1.007% 之间变化。所有模型的 MaxAPE 从最小值 1.708% 到最大值 4.765% 之间变化。因此，在各个模型的预报中平均不确定性大约为 1%，最大不确定性在 2.5% 左右（MaxAPE 的平均值），以及一个大约 4.8% 的最坏情况推测。这些值是否能代表一个可接受的不确定性的量取决于愿意去接受的风险的程度。

表 17-6 模型统计数据

Model Statistics						
Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
Subscribers for Market 1-Model_1	0	.245	10.663	17	.874	0
Subscribers for Market 2-Model_2	0	.245	35.583	17	.005	0
Subscribers for Market 3-Model_3	0	.244	15.787	17	.539	0
Subscribers for Market 4-Model_4	0	.469	19.126	15	.208	0
Subscribers for Market 5-Model_5	0	.572	11.574	16	.773	0
Subscribers for Market 6-Model_6	0	.254	10.982	17	.858	0
Subscribers for Market 7-Model_7	0	.005	33.268	17	.010	0
Subscribers for Market 8-Model_8	0	1.086E-15	10.836	18	.901	0
Subscribers for Market 9-Model_9	0	5.463E-5	35.018	17	.006	0
Subscribers for Market 10-Model_10	0	.330	14.768	17	.612	0

表 17-6 列出的是模型统计数据，从左向右各列的表头依次是模型名称、预测因子的数量、模型拟合统计（平稳  $R^2$ ）、Ljung-Box Q(18) 统计量、自由度、显著性概率值、异常值数量。第一列依次列出了 85 个供货商用户数的模型，表中是行列对应的计算结果。由于本表很大，限于篇幅，只列出了前 9 个模型的计算结果。

表 17-7 预测部分结果

Forecast				
Model		Jan 2004	Feb 2004	Mar 2004
Subscribers for Market 1-Model_1	Forecast	11503	11447	11390
	UCL	11686	11767	11870
	LCL	11321	11126	10910
Subscribers for Market 2-Model_2	Forecast	54893	55856	56704
	UCL	55632	57195	58575
	LCL	54154	54518	54832

YEAR_	MONTH_	DATE_	Predicted_Ma_rket_1_Model_1	Predicted_Ma_rket_2_Model_2	Predicted_Ma_rket_3_Model_3	Predicted_Ma_rket_4_Model_4
58	2003	10 OCT 2003	11820	51084	60880	171
59	2003	11 NOV 2003	11857	51273	60962	171
60	2003	12 DEC 2003	11687	53082	60516	181
61	2004	1 JAN 2004	11503	54093	56556	181
62	2004	2 FEB 2004	11447	55856	59305	181
63	2004	3 MAR 2004	11390	56704	58954	181

图 17-23 数据编辑器中的新变量

表 17-7 显示的是前两个模型（85 个模型）在指定的预测期 2004 年 1—3 月的预测值。UCL 和 LCL 分别为预测值置信区间（系统默认 95%）上限和下限。

图 17-23 给出的是数据窗中根据预测模型产生的预测值新变量。每个新变量包含估计期的模型预测值（1999 年 1 月到 2003 年 12 月），还包括指定的预测期 2004 年 1—3 月的预测值，因此增加了 3 个新样品。可以根据估计期的预测值观察模型拟合的优劣。

每个供应商的模型都有一个新变量，共 85 个；图中只显示了 4 个。

## 17.4 应用时间序列模型

应用时间序列模型程序从一个外部文件中取出存在的时间序列模型并应用它们到当前工作的数据文件。这里，新的或修改的数据是有效的，不用重新建立模型，读者可以使用本程序为时间序列获得预报。使用 Time Series Modeler 程序产生模型。

### 17.4.1 应用时间序列模型过程

按 Analyze→Time Series→Apply models 顺序单击菜单项，展开如图 17-24 所示的 Models 选项卡。

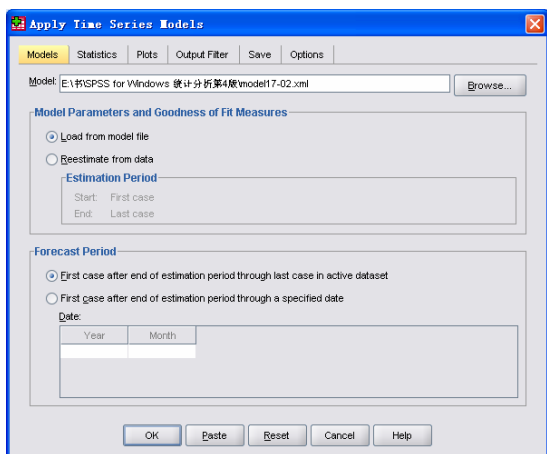


图 17-24 Apply Time Series Models 对话框

当拟合新模型或更新模型时，拟合优度反映使用数据的拟合情况。用本选项，对在当前工作的数据集中的因变量或自变量的预测都不考虑历史数据。如果想用历史的数据去影响预测，选择 Reestimate from data。另外，预测不考虑预测期中因变量的值，但考虑预测期中自变量的值。如果有更多的因变量序列的现值，并想要在预测中包括它们，就要重新估计并调整估计期包含这些值。

② Reestimate from data，根据当前工作的数据集中的数据重新估计模型参数。模型参数的重新估计不改变模型结构。例如，ARIMA(1,0,1)模型依旧，但要重新估计自回归和移动平均参数。不进行异常值的重新侦查。如果有异常值的话，那是来自模型文件的。

Estimation Period 栏显示的是默认的估计期。即重新估计模型参数的样品集。默认估计期包括当前工作数据集中的所有样品。可以使用 data 菜单 Select Cases 功能中的 Based on time 或 case range 重新定义估计期。估计期取决于有效数据，被程序使用的估计期可以通过模型改变，因此与显示值不同。对一个给定的模型，真实的估计期是在消除任何

#### 1. 打开 XML 文件

单击 Browse 按钮，找到由 Time Series Modeler 程序产生的 XML 模型文件。如 model17-02.xml。单击 Open 按钮。

#### 2. 选择模型参数和最佳拟合测度

在 Model Parameters and Goodness of Fit Measures 栏中，提供了两种模型参数和最佳拟合测度的选项，分别是：

① Load from model file，从模型文件中加载。使用文件中的模型参数产生预测。在输出中显示并用来过滤模型（最佳或最糟拟合）。拟合优度来自模型文件

邻近的缺失值后留下的部分，对模型的因变量，指定估计期的开始和结束。

3. Forecast Period 栏中定义预测期。具体内容参见 17.3.7 节中的内容。

4. 其他如 Statistics、Plots、Output Filter、Save、Options 选项卡中的定义参见 17.3.3 至 17.3.7 节中的内容。

## 17.4.2 应用时间序列模型分析实例

【例 4】以 17.3 中建立的 model17-02.xml 为模型基础，用在 data17-02.sav 基础上补充进 2004 年 1 月至 3 月各供货商用户数后形成的新数据集 data17-03.sav 进行预测 2004 年 4 月至 6 月的各月用户数，依次来说明 Apply models 的使用方法。

在 SPSS 中，进行 Apply models 的基本步骤如下：

1. 在 SPSS 数据编辑窗中，打开 data17-03.sav 文件。

2. 按 Analyze→Time Series→Apply models 顺序，展开 Apply Time Series Models 对话框，见图 17-24。

单击 Browse 按钮，然后选择数据盘中的 model17-02.xml。

为使时间序列的新值加入到预报中，Apply Time Series Models 程序必须重新估计模型参数。故选择 Reestimate from data 选项。由于模型的结构仍然是一样的，因此计算重新估计的时间远远快于原先建模的计算时间。

用来重新估计的样品集需要包括新数据。如果使用 First Case 到 Last Case 的默认的估计期，则这将是保险的。如果有时需要对除了系统默认外的事情设置估计期，则可以通过在 Select Cases 对话框中选择 Based on time 或 case range 来完成。

在 Forecast Period 中，选择 First case after end of estimation period through a specified date 选项。

在 Date 格中，在 year 框中输入 2004，在 month 框中输入 6。

数据集包含 1999 年 1 月至 2004 年 3 月的数据。用当前的设置，预测期将是 2004 年 4 月到 2004 年 6 月。

3. 单击 Save 按钮，在 Save 对话框的 Save 列中，选择（单击）Predicted Values，并用系统默认值 Predicted 作为 Variable Name Prefix。

模型预测值将被作为新变量存储在当前工作的数据集中，新变量使用前缀 Predicted。

4. 单击 Plots 按钮，由于我们对存储预测值作为新变量比产生预测图更感兴趣，所以在 Plots 对话框的 Plots for Individual Models 项中，撤销 Series 选项。这可以阻止为每个模型产生序列图。

5. 单击 OK 按钮运行，在输出窗中得到同表 17-4 模型描述、表 17-5 模型拟合、表 17-6 模型统计数据类似的结果，以及在当前工作文件中得到同图 17-23 数据编辑器中的新变量相类似的结果。有关它们的解释，参见 17.3.8 节中的相关图、表的解释。

## 17.5 自 相 关

自相关系数值度量了间隔不同的观察值之间的相关程度。它常用来洞悉产生数据的概率模型。解释自相关系数值集合的一个有效工具是相关图。相关图包括自相关函数（ACF）图和偏自相关函数（PACF）图两种。

使用 Autocorrelations 程序可以绘制 ACF 图及一个或一个以上序列的 PACF 图。需要注意：Autocorrelations 只适合于时间序列数据。

### 17.5.1 自相关图

要解释自相关图的含义是很困难的一件事情，这里只给出一般的基本的概念。

#### 1. 随机序列

如果时间序列是完全随机的，则当时间序列的长度  $N$  很大时，此时得到的自相关系数值近似服从均数为 0，方差为  $1/N$  的标准正态分布。根据置信区间的理论可知，在自相关图上，95% 的自相关系数值应出现在  $\pm 1.96 \cdot N^{-1/2}$  之间，这就是说，每 20 个自相关系数值至少应有 19 个自相关系数值应位于这个区间内，只有一个可能例外。但当随机序列中有异常值存在时，很可能看到不只一个自相关系数值出现在该区间之外。当时间序列中存在趋势或季节效应时，也能看到这种情况。这就要求我们在作自相关分析之前，要用以上所讲到的内容首先处理时间序列中的异常值，并从专业的角度来预判时间序列中是否存在趋势和季节效应，否则将给我们的分析带来巨大的麻烦。

#### 2. 短期相关

平稳序列常显示出短期相关，其明显的特征是，第一个自相关系数值很大，其后的自相关系数值大于 0 但逐渐减小，较长滞后的相关系数近似趋向于零。对这种类型的时间序列可用自回归模型来加以拟合。

#### 3. 非平稳序列

如果时间序列含有趋势，除非滞后值很大，否则自相关系数值是不会下降为 0 的。这是由于趋势的存在使得总均值一侧的观察值有大量后续的观察值也倾向于在均值的同一侧。对此类型的相关图，因为趋势支配所有其他特征，所以很难推出什么结论。因此在计算自相关系数值之前，要用前面提到的差分等手段消除时间序列中的趋势。

#### 4. 季节波动

含有季节波动的时间序列在自相关图上表现为出现相同频率的振荡。对于逐月观测的时间序列来说，第 6 个自相关系数值将是绝对值大而本身是负的，第 12 个自相关系数值将是大而正的。当自相关系数值形成正弦模型时，也会出现同样的规律。含有季节波动的数据，在数据的时序图上有十分清晰的表现。但对这种类型的季节数据，相关图没有提供更多的额外信息。如果将季节变化从数据中去除，如用各点的时间序列数据减去各个

季节对应点上的平均值后得到的消除季节影响后序列的自相关图,则可以进行分析了。

### 5. 交错序列

如果交错趋势存在时间序列中,则相关图也会出现交错的趋势。例如,第一个自相关系数值为负,则第二个自相关系数值为正。这是由于相邻的观察值总是总均值的两侧所造成的,这样凡滞后 2 的观察值总在总均值的同一侧,故第二个自相关系数值总是大于 0 的。

解释自相关图需要大量的实际工作经验,多学、多看、多做是掌握自相关图解释的唯一途径。至于各种时间序列的具体的特征图,不是本书的重点,在一般的时间序列分析书中都有介绍,在此不一一列出。

## 17.5.2 自相关分析过程

按 Analyze→Time Series→Autocorrelations 顺序,展开如图 17-25 所示的 Autocorrelations 对话框。

### 1. 定义变量

在源变量表中,选择一个或多个数值型变量,送入 Variables 框中。

### 2. 在 Display 栏定义显示函数:

① Autocorrelations 自相关。序列同滞后 1 或多个样品值的相关值。选择本项,计算 1、2、…、直到一个指定数的滞后的自相关。

② Partial autocorrelations 偏自相关,计算在干涉滞后相关的影响被消除之后,序列同滞后 1 个或多个样品值的相关值。

显示偏自相关需要解方程组,方程组的规模随滞后数的增大而增大。对高阶滞后(大于 24)要求做偏自相关要小心。即使在高速的计算机上,它也会比求自相关要花更长的时间。假如有季节因素影响的序列需要看看高阶滞后,则在确信序列是平稳序列之前建议你先看看自相关。然后,再要求做偏自相关。

### 3. 在 Transform 栏选择数据转换方法

① Natural log transform 自然对数转换。

② Difference 差分转换。选择此项需要在右边的文本框中输入正整数的差分值。

③ Seasonally difference 季节差分转换。在右边的文本框中输入正整数的差分值。当前周期显示在 Current Periodicity 后面。

### 4. 单击 Options 按钮,打开如图 17-26 所示的 Options 对话框。

(1) 在 Maximum Number of Lags 最大滞后数,默认值 16。可以重新输入最大滞后数。

(2) 在 Standard Error Method 栏选择计算标准误差的方法。如果在主对话框中,撤销

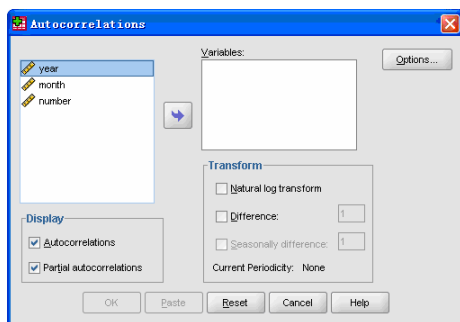


图 17-25 Autocorrelations 对话框



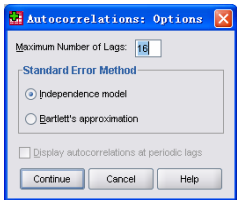


图 17-26 Options 对话框

Autocorrelations 选择，则这些选项无效。

① Independence model，独立模型，假设潜在过程是白噪声时的标准误差。

② Bartlett's approximation，Bartlett 逼近，用近似值计算标准误差，适用于序列描述  $k-1$  阶移动平均过程。用这种方法，标准误差随滞后的增加变大。

Display autocorrelations at periodic lags 显示周期性滞后处的自相关，如果已经定义了季节性，可以选择本项。

单击 Continue 按钮，返回主对话框。在主对话框单击 OK 按钮，运行本过程。

17.5.3 自相关分析实例

【例 5】数据 data17-04 中变量 sales 为某公司 1986—1997 年间各季度某商品的销售量数据，用自相关法对其进行统计学分析。打开 data17-04.sav 文件，分析步骤如下：

- 1. 按 Analyze→Time Series→Autocorrelation 顺序，展开 Autocorrelation 对话框。
- 2. 在源变量表中，选择销售量 Sales 作为分析变量，送入 Variables 框中。
- 3. 其他保持系统默认选项，单击 OK 按钮，运行，在输出窗中，出现表 17-8 至表 17-11 及图 17-27、图 17-28 所示的结果。
- 4. 结果解释

表 17-8 列出了模型的基本描述，从上到下依次为：模型名称 (MOD\_1)、序列名称 1 (销售量)、转换 (无)、非季节差分 (0)、季节差分 (0)、季节周期长度 (4)、滞后的最大阶数 (16)、关于计算的自相关标准误差的过程假设 (独立 (白噪声))、显示图形 (所有滞后)。

表 17-8 模型描述

Model Description	
Model Name	MOD_1
Series Name	销售量
Transformation	None
Non-Seasonal Differencing	0
Seasonal Differencing	0
Length of Seasonal Period	4
Maximum Number of Lags	16
Process Assumed for Calculating the Standard Errors of the Autocorrelations	Independence(white noise)
Display and Plot	All lags

Applying the model specifications from MOD\_1  
a. Not applicable for calculating the standard errors of the partial autocorrelations.

表 17-9 样品处理摘要

Case Processing Summary		销售量
Series Length		48
Number of Missing Values	User-Missing	0
	System-Missing	0
Number of Valid Values		48
Number of Computable First Lags		47

表 17-9 显示了样品处理摘要，从上到下分别为：序列长度 48，缺失值数量中用户缺失值 0、系统缺失值 0，有效值的数量 48，计算 1 阶滞后的数量 47。

表 17-10 显示的是自相关计算结果，从左向右，依次列出的是：滞后数、自相关系数数值、标准误差、Box-ljung 统计量 (值、自由度、原假设成立的概率值)。由于原假设 (假设基本过程是独立的，也即假定时间序列所反映的随机过程是白噪声) 成立的概

率值都小于 0.05，所以全部自相关均有显著性意义。

表 17-10 自相关计算结果

Autocorrelations					
Series: 销售量					
Lag	Autocorrelation	Std. Error <sup>a</sup>	Box-Ljung Statistic		
			Value	df	Sig. <sup>b</sup>
1	.432	.140	9.539	1	.002
2	.188	.138	11.377	2	.003
3	.317	.137	16.729	3	.001
4	.799	.135	51.565	4	.000
5	.288	.134	56.207	5	.000
6	.068	.132	56.471	6	.000
7	.164	.131	58.038	7	.000
8	.598	.129	79.517	8	.000
9	.143	.127	80.778	9	.000
10	-.049	.126	80.927	10	.000
11	.047	.124	81.070	11	.000
12	.451	.122	94.650	12	.000
13	.061	.121	94.902	13	.000
14	-.111	.119	95.773	14	.000
15	-.041	.117	95.896	15	.000
16	.318	.115	103.459	16	.000

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

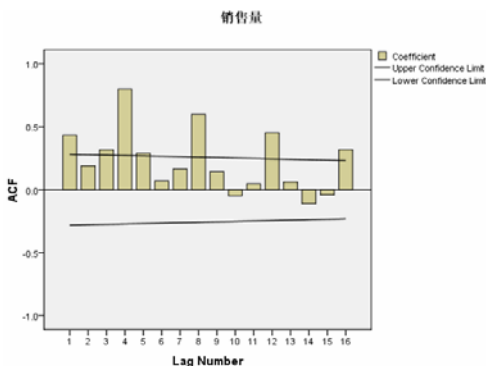


图 17-27 自相关图

表 17-11 偏自相关表

Partial Autocorrelations		
Series: 销售量		
Lag	Partial Autocorrelation	Std. Error
1	.432	.144
2	.001	.144
3	.289	.144
4	.752	.144
5	-.592	.144
6	.127	.144
7	-.051	.144
8	-.034	.144
9	-.064	.144
10	-.010	.144
11	.069	.144
12	-.039	.144
13	.057	.144
14	-.054	.144
15	-.076	.144
16	-.035	.144

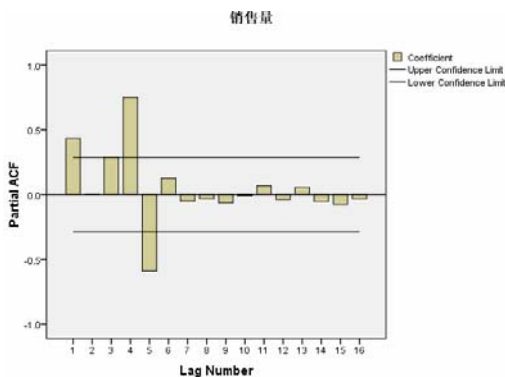


图 17-28 偏自相关图

表 17-11 从左向右显示的依次列出的是：滞后数、偏自相关、标准误差的计算结果。

图 17-27 显示的是对应于表 17-10 自相关系数值的自相关图。图 17-28 显示的是对应于表 17-11 偏自相关系数值的偏自相关图。

在滞后 4 处的重要顶点暗示在数据中存在周期为 4（4 个季度）的季节成分。检查偏自相关函数图同样可得到这个十分明确的结论。

## 17.6 季节分解法

在实际工作中，经常会遇到按日、周、月、季或年记录的数据资料，如每天新生儿出生的情况、某产品每月的销售量、每年 GDP 的增长率等。这些资料通过自相关分析可能符合季节性分布，对这些有随机变异、长期趋势、季节效应或周期变动的时间序列资

料, 可以使用季节分解法对其进行分析, 从而得到有意义的结果。

### 17.6.1 季节分解法分析过程

1. 进行季节分解的数据, 要有至少包括 4 个完整季节数据的变量。打开 Data 菜单中的 Define Dates 对话框, 定义时间序列的周期。定义了周期后才能进行季节分解。

2. 按 Analyze→Time series→Seasonal Decomposition 顺序展开季节分解对话框, 见图 17-29。本程序用来估计时间序列的乘性或加性季节因素。

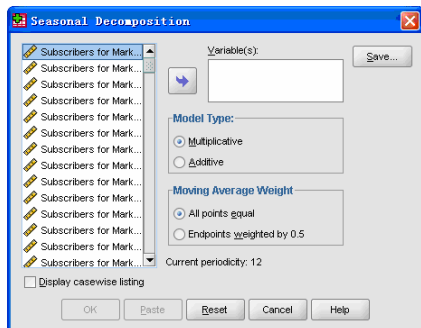


图 17-29 季节分解主对话框

3. 指定需要季节分解处理的变量。从源变量表中选择分析的变量, 移到 Variable(s)框中。该变量必须包括 4 个完整的季节数据。

4. 在 Model 栏中, 根据时间序列构成的特点, 选择 Multiplicative 乘法模型或 Additive 加法模型。

5. 在 Moving Average Weight 移动平均的权重栏中, 指定在计算移动平均时如何对待时间序列:

① All points equal, 计算周期跨度相等和所有点权重相等时的移动平均, 常用于周期是奇数的情形。

② Endpoints weighted by 0.5, 用相同跨度 (周期+1) 和端点权重乘 0.5 计算移动平均。这个选项仅当时间序列的周期是偶数时有效。

6. Display casewise listing, 在运算过程中对每个变量生成一行 4 个新序列值。

7. 单击 Save 按钮, 展开保存对话框, 如图 17-30 所示。

① Add to file, 季节分解产生的新序列作为新变量彼此在数据窗中。变量名由三部分组成: 三个字母的前缀、下划线、数字。这是系统默认的保存方法。

② Replace existing, 季节分解产生的新序列作为临时变量彼此在数据窗中, 已经存在的临时变量被剔除。变量名由三部分构成: 三个字母的前缀、井号和一位数字。

③ Do not create, 新序列不添加到数据文件中。

8. 单击 OK 按钮, 系统立即执行命令。

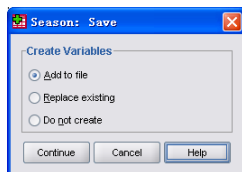


图 17-30 Save 对话框

### 17.6.2 季节分解法分析实例

【例 6】数据 data17-04 中变量 sales 为某公司 1986—1997 年间各季度某商品的销售额数据, 用季节分解法对其进行统计学分析。

## 1. 操作方法

- (1) 按 Analyze→Time series→Seasonal Decomposition 顺序展开如图 17-25 的对话框。
- (2) 选择销售量 sales 变量进入 Variables 对话框。
- (3) 在 Model 框中, 选定 Multiplicative 项。
- (4) 在 Moving Average Weight 框中, 选定 All points equal 项。
- (5) 使用 Save 对话框中默认设置。
- (6) 单击 OK 按钮, 执行运算。

## 2. 输出结果, 见表 17-12、表 17-13。

表 17-12 给出了模型描述, 从上到下依次为: 模型名称: MOD\_4、模型类型: 乘法模型、序列名 1: 销售量、季节周期长度: 4、移动平均计算方法: 跨度等于周期且所有点权重相等。

表 17-13 列出了季节及其对应的季节因素指数。

在数据窗口中生成来自给定模型的销售量的误差项 (ERR\_1)、季节校准序列 (SAS\_1)、季节因素指数 (SAF\_1)、季节趋势周期 (STC\_1) 4 列新数据, 见图 17-31。

表 17-12 模型描述

Model Description	
Model Name	MOD_4
Model Type	Multiplicative
Series Name 1	销售量
Length of Seasonal Period	4
Computing Method of Moving Averages	Span equal to the periodicity and all points weighted equally

Applying the model specifications from MOD\_4

表 17-13 季节因素

Seasonal Factors	
Series Name:销售量	
Period	Seasonal Factor (%)
1	111.8
2	109.2
3	75.8
4	103.2

	sales	year	quarter	date	ERR_1	SAS_1	SAF_1	STC_1
1	3017.60	1986	1 Q1 1986		0.98512	2698.66383	1.11818	2739.42113
2	3043.54	1986	2 Q2 1986		1.01355	2787.17429	1.09198	2749.92652
3	2094.35	1986	3 Q3 1986		0.99748	2763.94145	0.75774	2770.93731
4	2009.84	1986	4 Q4 1986		0.97090	2722.45957	1.03210	2804.06932
5	3274.80	1987	1 Q1 1987		1.03047	2928.67985	1.11818	2842.09329
6	3163.28	1987	2 Q2 1987		1.01125	2896.82826	1.09198	2864.59796
7	2114.31	1987	3 Q3 1987		0.96847	2790.26293	0.75774	2881.11122

图 17-31 数据文件中增加的 4 个新变量

# 17.7 频谱分析

## 17.7.1 频谱分析概述

当把时间序列看作是由不同频率的正弦、余弦波组成时, 就可用 Schuster 在 1898 年引入的周期图来进行时间序列分析了。周期图最初是用来检测和估计混在噪声中、频率为已知的正弦分量的振幅。用它提供的方法也可检验序列的随机性。在周期图的基础上, 用功率谱, 可以建立样本谱。同样, 样本谱也可用来检验和估计隐含于噪声中未知频率的正弦分量的振幅。尤其当我们事先已知频率  $f$  并不具有与序列长度的谐振关系时, 样本谱更是实现上述目的的有力工具。可以证明, 样本谱是自协方差函数估计值的傅里叶余弦变换。这为谱分析理论奠定了基础。

当平稳时间序列的频率、振幅和相位都是随机变化时, 样本谱失去了其应有的作用, 此时用频率强度的均值建立起来的谱分析是最重要的分析工具。

Spectral Plots 频谱图程序可用来识别时间序列中的周期行为。而不是分析一个时间点向下一个时间点的变化, 通常是把分析序列的变化转化成不同频率的周期成分。平稳序

列在低频时有更强的周期成分,随机变化的白噪声遍及所有频率。分析变量应该是数字型、平稳的不包含缺失值的时间序列。应从时间序列中减去任何的非零均值。应在预测分析前处理缺失值,方法参见替换缺失值方面的内容。将不稳定序列变成平稳序列的常用方法是差分转换。参阅建立时间序列方面的内容。

## 17.7.2 频谱分析过程

1. 按 Analyze→Time series→Spectral Analyze 顺序展开频谱分析对话框,见图 17-32。
2. 在源变量表中,选择一个或多个数值变量,送入 Variables 下框中。
3. 在 Spectral Windows 的下拉列表中,选择平滑序列的滤波算法获取谱密度估计。

(1) Tukey-Hamming 法权重公式为

$$W_k = 0.54D_p(2\pi f_k) + 0.23D_p(2\pi f_k + \frac{\pi}{p}) + 0.23D_p(2\pi f_k - \frac{\pi}{p}) \quad k = 0, L, p$$

式中,  $p$  是跨度一半的整数部分,  $D_p$  是  $p$  阶的 Dirichlet 中心。

(2) Tukey 法权重公式为

$$W_k = 0.5D_p(2\pi f_k) + 0.25D_p(2\pi f_k + \frac{\pi}{p}) + 0.25D_p(2\pi f_k - \frac{\pi}{p}), \quad k = 0, L, p$$

式中,  $p$  是跨度一半的整数部分,  $D_p$  是  $p$  阶的 Dirichlet 中心。

(3) Parzen 法权重公式为

$$W_k = \frac{1}{p}(2 + \cos(2\pi f_k))(F_{p/2}(2\pi f_k))^2, \quad k = 0, L, p$$

式中,  $p$  是跨度一半的整数部分,  $F_{p/2}$  是  $p/2$  阶的 Fejer 中心。

(4) Bartlett 法计算上半部分谱窗的权重为

$$W_k = F_p(2\pi f_k), k = 0, L, p$$

式中,  $p$  是跨度一半的整数部分,  $F_p$  是  $p$  阶的 Fejer 中心。下半部分同上半部分对称。

(5) Daniell (Unit) 法计算谱窗形状的权重都等于 1。

(6) None, 不用做滤波处理。谱密度估计同周期图相同。

4. 在 Span 后的框中指定跨度。即横跨执行平滑的连续值的范围。通常使用奇整数。平滑谱密度图多使用大跨度, 较少使用小跨度。系统默认值为 5。

5. 中心化变量的选择

Center variables, 在计算频谱前, 校准序列使

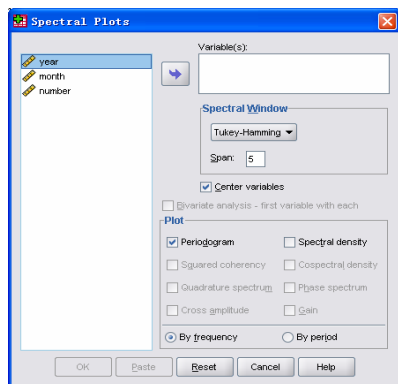


图 17-32 谱图分析对话框

其有 0 均数（中心化），并剔除同序列均数有关联的大量的项（剔除异常值）。因为谱分析时，相应序列平均数频率为 0，否则周期图没什么实际意义。为此先使数据以 0 为中心。

**Bivariate analysis**，双变量分析。如果选择两个或两个以上分析变量，可以选择本选项，要求做 **Variable(s)**框中第一个变量（因变量）与后面每一个变量（自变量）的双变量谱分析。各个序列的单变量分析照样进行。

#### 6. 在 Plot 栏中选择输出的分析图

周期图和谐密度图对单变量和双变量分析都有效。其他选项只对双变量分析有效。

(1) **Periodogram** 周期图，以频率或周期为横轴的非平滑的频谱振幅图（在对数标尺上绘制）。变异均匀地分布在所有波段象征“白噪声”。

(2) **Squared coherency**，两序列提衰量的乘积。

(3) **Quadrature spectrum**，交叉周期图的虚部，它是两个时间序列异相频率分量相关的测度。分量是  $\pi/2$  弧度的异相。

(4) **Cross amplitude**，余谱密度平方与正交谱平方之和的平方根。反映振幅的大小。

(5) **Spectral density**，已过滤去不规则变化的周期图。

(6) **Cospectral density** 交叉周期图的实部，是两个时间序列同相频率分量相关的测度。

(7) **Phase spectrum**，一个序列领先或滞后其他序列的各频率分量的长度的测度。

(8) **Gain**，用谱密度为序列之一划分的交叉振幅的商。两个序列中每一个都有其自己的提衰量值。它是在某一频率下的回归系数，同线性回归系数类似。

**By frequency**，所有图都由频率生成，频率的范围在频率 0（常数项或均数项）到频率 0.5（两个观察资料的周期项）之间。

**By period** 选项，所有图都由周期生成，周期的范围在周期 2（两个观察资料的周期项）到周期等于观察值的数量（常数项或均数项）之间。周期在对数标尺上显示。

### 17.7.3 频谱分析实例

【例 7】Data17-05.sav 中记录的是国际航线 1949 年 1 月至 1960 年 12 月间月度旅客总数（单位：千人），试用频谱分析法分析其是否有年度周期。

#### 1. 打开数据文件后的操作步骤

(1) 单击 **Analyze**→**Time series**→**Spectral**

**Analyze** 展开如图 17-31 谱图对话框。

(2) 选择 **number** 送入 **Variables** 框中。在 **Plot** 项中选择 **Spectral density**。

(3) 单击 **OK** 按钮，执行运算。

#### 2. 输出结果，见表 17-14、图 17-33、

图 17-34。

表 17-14 模型描述

Model Description	
Model Name	MOD_12
Analysis Type	Univariate
Series Name	number
Range of Values	Reduced by Centering at Zero
Periodogram Smoothing	Tukey-Hamming
Spectral Window	
Window Span	5
Weight Value	WK(2)
	2.233
	WK(1)
	2.238
	WK(0)
	2.240
	WK(1)
	2.238
	WK(2)
	2.233

Applying the model specifications from MOD\_12

表 17-14 给出了模型的描述，从上到下依

次是：模型名称（MOD\_12）、分析类型（单变量）、序列名 1（number）、值范围（通过中心在 0 点处理）、周期图平滑：谱窗口（Tukey-Hamming）、窗口跨度（5）、权重值： $w(-2)=2.233$ 、 $w(-1)=2.238$ 、 $w(0)=2.240$ 、 $w(1)=2.238$ 、 $w(2)=2.233$ 。

图 17-33 是周期图，周期图中显示的背景噪声中，有引人注目的连续的峰值，在小于 0.1 的最低频率处有最高的峰值，因此可以怀疑数据中包含一个年度的周期成分，年度成分的贡献组成了周期图。在时间序列中每个数据点表示一个月，因此一个年度周期对应于当前数据集中的周期 12。由于周期和频率互为倒数，周期 12 对应  $1/12$ （或 0.083）的频率。所以年度成分暗示在周期图中 0.083 处的一个峰值，它与正好低于 0.1 的频率处出现的峰值相一致。

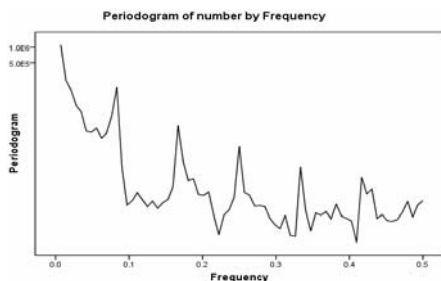


图 17-33 周期图

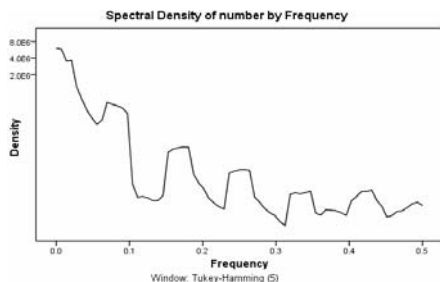


图 17-34 密度图

图 17-34 显示的是谱密度图。是经过消除背景噪声平滑后的周期图。残余峰值最好同谱密度函数一起分析，让潜在结构变得更加清楚独立。谱密度由 5 个明显的等间隔出现的峰值组成。最低频率峰值是在 0.0833 处。分析变量时间序列可以分解成 4 个主要（幅度较大）正弦或余弦成分。它们的周期即峰值点频率的倒数。

## 17.8 互 相 关

### 17.8.1 互相关概述

ACF 和 PACF 是描述单个时间序列的重要工具。但在很多场合下，需要考虑的时间序列不是一个，而是同时需要考虑多个时间序列之间的关系。例如，市场的货币供应量和股价变化之间的关系、某产品的广告投入和该产品市场占有率及销售量的关系。这时就需要考虑两个序列或多个序列之间的相互关系。为了和单序列分析（也称单变量时间序列分析）区分，将这里将讨论的问题的模型称为多序列分析（或多元时间序列分析）。分析这种模型的工具是互相关函数。

所谓互相关函数（CCF）是指两个时间序列间的相关。即一个序列的观察值同另一个序列在不同的滞后和领先时的观察值之间的相关关系。互相关通常显示在图中，称为

互相关图。互相关图可以帮助识别那些是其他变量先行指数的变量。

### 17.8.2 互相关过程

1. 本程序用来为正、负和 0 阶滞后绘制两个或多个序列互相关函数图。互相关程序只适用于时间序列数据。

2. 按 Analyze→Time series→Cross-Autocorrelation 顺序展开如图 17-35 互相关分析主对话框。

3. 在源变量表中,至少选择两个变量,送入 Variables 下框中。

4. 在 Transform 中定义序列的转换方法: Natural log transform、Difference、Seasonally difference 三个函数的说明可参阅 17.2.1 节中的相关内容。

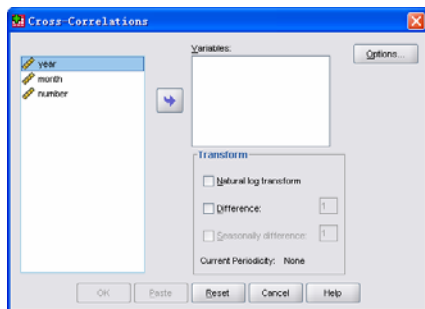


图 17-35 Cross-Correlations 对话框

在这些选项下面的 Current periodicity:后显示当前的周期。

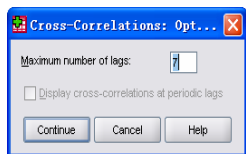


图 17-36 选项对话框

5. 单击 Options 按钮,弹出 Options 对话框,见图 17-36。在 Maximum number of lags 框中输入互相关的最大滞后数,默认值为 7。如果数据中定义了季节,只显示选择项 Display cross-correlations at periodic lags 显示周期延迟处的互相关。

6. 单击 Continue 按钮,返回主对话框。单击 OK 按钮,系统立即执行命令。

### 17.8.3 互相关实例

【例 8】Data17-06.sav 中记录的 1989 年 1 月至 1998 年 12 月间某公司每月三种男、女服装产品的销售量情况,试对其进行互相关分析。

(1) 在数据编辑窗口中,打开 Data17-06.sav。按 Analyze→Time series→Cross-Autocorrelation 顺序展开如图 17-35 所示的互相关对话框。

(2) 选择男装销售额变量 men 和女装销售额 women 变量,送入 Variables 框中。

(3) 其他保持系统默认选择,单击 OK 按钮,执行运算。

输出结果,见表 17-15~表 17-17、图 17-37。

表 17-15 给出了模型的描述,从上到下依次是:模型名称(MOD\_1)、序列名 1(男子服装销售量)、序列名 2(女子服装销售量)、转换(无)、非季节差分(0)、季节差分(0)、季节周期的长度(无周期)、滞后的范围从-7 到 7。显示并绘图(全部滞后)。

表 17-16 是样品处理摘要,从上到下依次是:序列长度(120)、由于读者缺失值排除的样品的数量(0)、由于系统缺失值排除的样品的数量(0)、有效样品数量(120)、



表 17-15 模型描述

Model Description		
Model Name	MOD_1	
Series Name	1	Sales of Men's Clothing
	2	Sales of Women's Clothing
Transformation	None	
Non-Seasonal Differencing		0
Seasonal Differencing		0
Length of Seasonal Period	No periodicity	
Range of Lags	From	-7
	To	7
Display and Plot	All lags	
Applying the model specifications from MOD_1		

表 17-16 样品处理摘要

Case Processing Summary	
Series Length	120
Number of Excluded Cases Due to	User-Missing Value 0
	System-Missing Value 0
Number of Valid Cases	120
Number of Computable Zero-Zero Correlations After Differencing	120

在差分后计算 0 阶相关的数量（120）。

表 17-17 给出的是互相关系数的计算结果表，从左向右各列列出的依次是滞后、互相关系数值和互相关系数的标准误差。

图 17-37 给出的是男女服装销售量之间的互相关图，它用表 17-17 中滞后的值作为横坐标，用互相关系数值作为纵坐标，通过直方图的形式表现了出来。最大互相关系数出现在滞后 0 处，为 0.802，显然，互相关系数并不关于滞后 0 处对称。滞后 0 处的相关同简单的两个变量间的皮尔逊相关是一样的。说明两个变量之间存在线性相关。图 17-36 中的横轴上下的两根横线，分别是置信限的上、下限。

表 17-17 互相关系数表

Cross Correlations		
Series Pair Sales of Men's Clothing with Sales of Women's Clothing		
Lag	Cross Correlation	Std. Error <sup>a</sup>
-7	.159	.094
-6	.150	.094
-5	.211	.093
-4	.224	.093
-3	.271	.092
-2	.342	.092
-1	.374	.092
0	.802	.091
1	.134	.092
2	.114	.092
3	.125	.092
4	.209	.093
5	.163	.093
6	.124	.094
7	.178	.094

a. Based on the assumption that the series are not cross correlated and that one of the series is white noise.

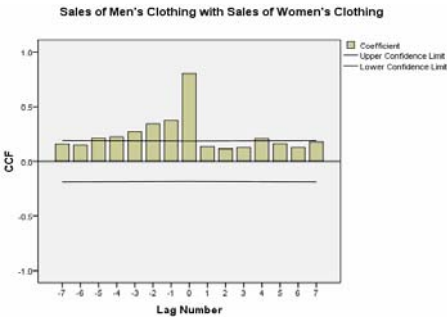


图 17-37 男女服装销售量的互相关图

习 题 17

1. 简述时间序列的基本概念。时间序列分析过程中有哪几种常用的方法？
2. 对数据用时间序列模型进行拟合处理前，应做哪些准备工作？
3. 在哪个过程中可进行缺失值的修补？修补缺失值的方法共有几种？
4. 在哪个过程中可定义时间变量？
5. 时间序列分析是建立在序列的平稳的条件上的，怎样判断序列是否平稳？
6. 为什么要建一个时间序列的新变量？在 SPSS 的哪个过程中建时间序列新变量？
7. 光盘中 Data17-07.sav 记录了一个邮购公司在 1989 年 1 月至 1998 年 12 月间男、女服装产品的销售量情况以及一些可能影响服装销售的宣传、服务方面的变量。试用学过的时间序列方法对其进行分析，并预测 1999 年 3 月的男装的销售量。

# 第 18 章 多响应变量的分析

## 18.1 多响应变量的概念与分类

### 1. 多响应变量的概念

SPSS 软件中大部分分析过程解决等间隔测度的数值型变量、分类变量或称分组变量（定序变量或名义变量）分析的问题。这些变量在每个观测中都有一个并且只有一个确定的值。在当前社会实践活动中大量存在这样的变量，对于一个确定的观测对象，该变量有几个值与之对应。例如，当问到您喜欢什么颜色时，您可能既喜欢红色，也喜欢蓝色和绿色。如果让您按喜欢程度排一下顺序时，您的回答是：红色第一，蓝色第二，绿色第三。这就构成了对一个问题（变量）的多个选择（响应）。这种问题称作多项选择题。目前，市场研究或许多领域对某事物评价的研究中常常遇到这样的问题。

### 2. 多响应变量的分类与代码

多响应变量的分类取决于对问题的设计和对数据的整理及其数据文件的建立。

#### (1) 多响应二分变量集及其编码

多响应二分变量集是由若干个二分变量组成的变量集。这些二分变量反映了一个问题的多个可能的答案。例如对下面的一组问题由 9 个问题组成，每个问题可选择的答案由一个变量表示，每个变量的值只能有表明“是”、和“否”的两个代码。

表 18-1 为向顾客发放的颜色调查表，在选择服装时您喜欢什么颜色作为主体颜色，在答案前的“□”中画“√”（可多选）。这是一组问题，每个问题均有两个答案，回答者只能选择其中一种。在建立数据文件时，变量名使用相同的变量主名，后面加以不同序号组成，本组问题的 9 个变量名是 color1~color9，以便分析和整理时识别。答案的编码规则为：回答“是”变量值为 1，回答“否”变量值是 0，其他值为缺失值。

对这 9 个问题需要放在一起分析，就要组成变量集，称作多响应二分变量集。对多响应二分变量集可以进行频数分布分析和与其他分类变量做交叉表分析。

这样的问题设计的优点是每个回答者对每种颜色均可以表示她（他）的态度。问题明确，回答时可以很少考虑，答题迅速，答题的要求容易被接受。

表 18-1 服装颜色问卷

编号	调查内容	选 项	
1	您喜欢红色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
2	您喜欢橙色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
3	您喜欢黄色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
4	您喜欢绿色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
5	您喜欢青色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
6	您喜欢蓝色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
7	您喜欢紫色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
8	您喜欢黑色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
9	您喜欢白色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否

(2) 多响应分类变量集及其分析方法

多响应分类变量集是由若干个分类变量组成。每个分类变量都有两个以上的值作为回答者的答案。这些分类变量共同反映了回答者对问题的看法，因此单个分析就会有失全面。例如下面的问题及其对回答者的要求结果造成每个观测量都有 3 个可能的回答，使用一个变量不能完全概括 3 个答案，因此必须建立 3 个变量，每个变量均有 10 个可能的答案，即需要 10 个代码表示。

作为服装主体颜色，您可以选择最喜欢的 3 种，在答案前的○中填写喜欢的序号号（最喜欢的为①，其次的为②、③）。

- ☐ 红    ☐ 橙    ☐ 黄    ☐ 绿    ☐ 青
- ☐ 蓝    ☐ 紫    ☐ 黑    ☐ 白    ☐ 说不清

这是一个问题，每个问题可以有 3 个答案。在建立数据文件时，要建立 3 个变量，color1~color3 表示答者按喜欢程度选择的 3 个颜色。答案变量的值均按填写的顺序值编码。即代码 A 表示选择红色、代码 B 表示选择橙色、C 表示选择黄色……示选择白色、J 表示说不清。例如，选择结果为①黑、②红、③蓝，则变量 color1 的值为 H，变量 color2 的值为 A，变量 color3 的值为 F。当然也可以使用数字编码。

(3) 解决多响应问题的 SPSS 过程

无论哪种多项选择题，由于每个大问题包含若干个子问题。在分析时如果使用单个变量进行分析肯定是不全面的，因此在 SPSS 中首先将每个题的若干答案组成一个综合变量即变量集（Set），然后对综合变量的各种取值进行分析。

多项选择问题的分析在 SPSS 中是通过 Analyze 的菜单项 Multiple Response 中的各项功能实现的，见图 18-1。

① Define Sets 定义并建立多响应二分变量集或多响应分类变量集。

② Frequencies 对多响应二分变量集和多响应分类变量集进行频数分布分析。

③ Crosstabs 对多响应二分变量集、多响应分类变量集与其他变量集或原变量进行交叉

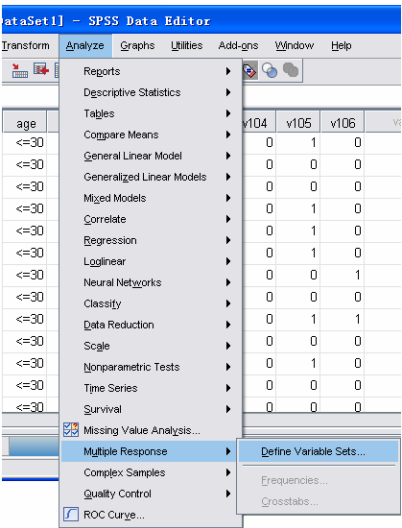


图 18-1 完成多响应问题的菜单

表分析。

对多响应的二分变量集或多响应分类变量集也可以使用表格功能矩形分析。即 Analyze 菜单中的 Tables 命令，利用表格功能中的统计分析功能得到更丰富的统计量和统计分析结果。

## 18.2 定义与建立多响应变量集

### 1. 定义多响应变量集

定义多响应变量集是对多项选择题进行分析的必要步骤，必须把一组反映同一问题的多个答案变量组合在一个变量集中，方法如下：

(1) 按 Analyze→Multiple Response→Define Sets 顺序单击各菜单项，最后打开 Define Multiple Response Sets 定义多响应集对话框，如图 18-2 所示。

(2) 在 Set Definition 栏里选择同属于一个问题的多个答案变量，通过向右箭头按钮送入 Variable in Set 栏内。再根据该栏内的变量，定义变量集。

(3) 在 Variables Are Coded As 栏内定义这组变量的编码方式。

① Dichotomies Counted value，二分变量的计数值。如果所选择的变量是回答“是”、“否”的题目，选择此项，并在其后的编辑栏内输入想进行计数的答案代码。如要对回答“是”的观测进行计数，并且在数据文件中对每个问题选择“是”使用代码为 1，则在编辑栏内输入 1。

② Categories，分类变量。如果所选择的每个变量的回答是表示赞同顺序的数字，应该选择此项，并在其后的两个编辑栏内，输入要分析的变量的取值范围，即其值的起止范围。

(4) 在 Name 栏内为变量集命名。

(5) 在 Label 栏内输入变量集的标签。

(6) 单击 Add 按钮将定义好的变量名及其标签送入右面的 Multiple Response Sets 栏内。该栏在命名的多响应变量集前自动加“\$”以区别于一般变量。

(7) 反复上述操作，定义多个多响应变量集。单击 Close 按钮结束。

使用以上功能菜单和上述方法定义的多响应变量集，只能在图 18-1 中所示二级菜单的各项中使用也可以用在 Tabel 中构建自定义表格。按上述的方法定义的多响应变量集可以进行频数分布分析和交叉表分析。

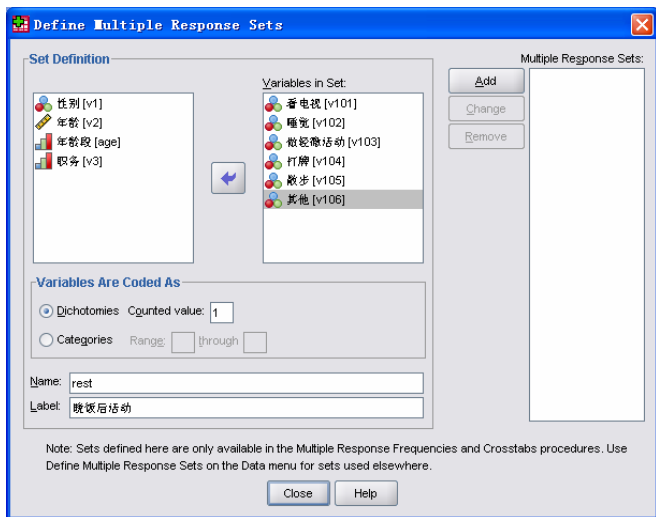


图 18-2 定义多变量集的对话框

## 18.3 多响应变量的频数分布分析

多响应变量集的频数分布分析操作简单,下面举例说明之。

一项对公务员的营养、运动及心理调查中,除记录了被访者的性别、年龄外,还有如下问题和供选择的答案:

问题:一般您在晚饭后做什么?(可多选)

供选答案:A)看电视,B)睡觉,C)轻微活动,D)打牌,E)散步,F)其他(如看电影、跳迪斯科、加班、应酬等)。

### 18.3.1 多响应二分变量集的频数分布分析

【例1】根据答案建立了6个变量V101~V106,选择的变量值代码为1,未选择为0。数据见data18-01。

按18.1节介绍的方法建立多响应变量集rest标签为“晚饭后活动”。

1. 使用Multiple Response的Frequencies进行频数分布分析的步骤如下:

(1) 按Analyze→Multiple Response→Frequencies顺序单击菜单项,打开多响应变量集的频数分布分析对话框,如图18-3所示。

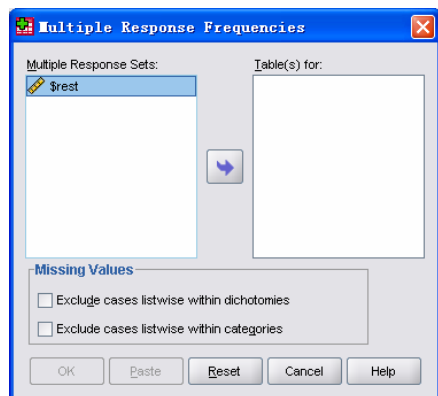


图 18-3 频数分布分析对话框

(2) 在Multiple Response Frequencies对话框中,已经定义的变量集显示在左面Multi Response Sets栏中。选择要进行频数分布分析的变量集,本例只有一个,选中并送入右面的Table(s) for栏中。

(3) 在Missing Values栏中选择处理缺失值的方法。

- Exclude cases listwise within dichotomies,

多响应二分变量集中任意一个变量值缺失的观测量从分析中剔除。只有当多响应变量集的所有组成变量是二分变量时才可以选择此项缺失值处理方法。系统默认的是只剔除多响应变量集中的所有

变量都没有计数值的观测量。也就是说观测量在多响应二分变量集中至少有一个变量包括计数值,该观测量就会计入频数分布表中。

- Exclude cases listwise within categories, 多响应分类变量集中任意一个变量值不在定义范围内,被认为是缺失的观测量从分析中剔除。只有当多响应变量集的所有组成变量是分类变量时才可选择此项缺失值处理方法。默认的是当一个观测量的组成多响应分类集的所有变量没有一个变量的值包括在定义的范围之内时,该观测量才被认为是缺失

的，要从分析中剔除。

显然，无论是二分变量集还是分类变量集，默认的处理方法，数据利用率较高。

## 2. 语句和语句说明

(1) 单击 Paste 按钮得到下列简单程序：

MULT RESPONSE

GROUPS=\$rest '晚饭后活动' (V101 V102 V103 V104 V105 V106 (1)) /FREQUENCIES=\$rest.

(2) 语句说明

MULT RESPONSE 语句调用 Multiple Response 过程。

GROUPS 语句指定分析多响应变量集名 rest，由 V101~V106 组成，每个变量的计数值是 1 的二分变量。

语句/FREQUENCIES=\$rest 要求对变量集 rest 进行频数分布分析。

## 3. 输出结果及说明，输出结果见图 18-2、18-3。

(1) 表 18-2 是观测量小结：合法观测量共 531 个，占 99.1%；缺失值观测量就是在二分变量集中的变量没有一个值是 1 的。都为 0 或者有 0、1 以外的值，这样的值有 5 个，占 0.9%。

表 18-2 观测量小结

Case Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
\$rest <sup>a</sup>	531	99.1%	5	.9%	536	100.0%

a. Dichotomy group tabulated at value 1.

表 18-3 多响应二分变量集的频数分布表

\$rest Frequencies				
		Responses		Percent of Cases
		N	Percent	
晚饭后活动	看电视	381	47.5%	71.8%
	睡觉	57	7.1%	10.7%
	做轻微活动	127	15.8%	23.9%
	打牌	18	2.2%	3.4%
	散步	175	21.8%	33.0%
	其他	44	5.5%	8.3%
Total		802	100.0%	151.0%

a. Dichotomy group tabulated at value 1.

(2) 表 18-3 是多响应二分变量集中各变量的频数分布表。

① 组合的多响应变量集名 \$rest，标签“晚饭后活动”。表中出现的都是变量标签。

② 标注 a 说明表中是计数值为 1 的频数。

③ Response 是各变量响应的计数 N 和占回答为 1 的总数的百分比。

④ 表头为 N 的列对应左边变量值为 1 的发生频数，其总和 802，因为允许多选所以大于观测量总数 536。（其中 5 个缺失值、531 个有效值）802 为总选择为 1 的答案数。

⑤ Percent 说明 N 中的频数占总答案数 802 的百分比。总百分比为 100%。

⑥ Percent of Cases 说明 N 中的频数占总观测量 536 的百分比。Total 相当于 801 占总观测量数 536 的百分比，因此大于 100%。

从频数分布表中可以看到，晚饭后看电视的比例，大大超过其他活动的比例。散步或做轻微活动的比例相对较少。这对健康是不利的，应该引起重视。

## 18.3.2 多响应分类变量集的频数分布分析

【例 2】使用与 18.3.1 节相同的例题，如果问题是：按照您的习惯选择 3 个晚饭后

的主要活动，并按经常性排列顺序。例如，最经常的是晚饭后看电视，其次是散步，有时打打牌，则应该在看电视前填写①，在散步前填写②，在打牌前填写③。

- 看电视
- 睡觉
- 轻微活动
- 打牌
- 散步
- 其他（继续工作、看电影、跳迪斯科或应酬等）

1. 建立数据文件，见 data18-02。变量 vv1~vv3 分别表示第一选择到第三选择。
2. 对这类问题获取的数据进行频数分布分析的方式有两种：

(1) 一种是对三个变量分别进行频数分布分析。

- ① 按 Analyze→Descriptive Statistics→Frequencies 顺序打开频数分布分析对话框。
- ② 在主对话框中，将 vv1、vv2、vv3 三个变量送入右面的 Variables 栏，并选择 Display frequency tables 选项，要求输出频数分布表。

③ 单击 Chart 按钮，在对话框的 Chart Type 栏中选择 Pie Chart 饼图，在 Chart Value 栏选择 Percentage 要求以百分比标注饼图的分块。

④ 程序语句及解释  
FREQUENCIES

VARIABLES=vv1 vv2 vv3 /PIECHART PERCENT /ORDER= ANALYSIS.

FREQUENCIES 命令语句调用 Frequencies 过程进行频数分布分析。  
VARIABLES 命令语句指定分析三个变量 vv1、vv2、vv3 的频数分布。  
PIECHART 命令语句要求做饼图，PERCENT 参数要求以百分比标注。  
ORDER 语句输出顺序是分析变量出现的顺序（即分析顺序）。

⑤ 输出结果见表 18-4~表 18-7 和图 18-4~图 18-6。

表 18-4 观测量统计表

表 18-4 是观测量小结，显示了每种选择的有效值和缺失值。

Statistics				
		第一选择	第二选择	第三选择
N	Valid	532	497	274
	Missing	4	39	262

从表 18-5 第一选择的频数分布表，和百分比饼图 18-4，可以一目了然地看出：

- 52.4% 的问卷回答者晚饭后活动的第一选择是看电视。是他们最经常的活动方式。
- 其次是散步 23.3% 和轻微活动 16.8%。可见电视在人们的晚饭后活动中占有很重要的地位。看电视是大多数人首选的活动。做轻微活动的人和散步的人总和占 40.1%，说明当前相当一部分人很重视晚饭后活动。
- 打牌、睡觉和其他活动（或许是看电影、跳迪斯科、加班工作或应酬等）的人数比例很少，总和不到 10%。

从第二选择的频数分布表和百分比饼图可以看出，第二选择看电视和散步的百分比相当，做轻微活动的占有较大比例。

第三选择的频数分布表和百分比饼图表示将近一半的问卷回答者没有任何选择。也就是说近 50% 的人每天晚饭后经常安排两项活动说明公务员业余生活比较单调。

如果想分析晚饭后看电视、散步等活动所占的总百分比就应该建立多响应分类变量

集，并对变量集进行频数分布分析。

表 18-5 第一选择频数分布表

第一选择		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	看电视	281	52.4	52.8	52.8
	睡觉	11	2.1	2.1	54.9
	轻微活动	90	16.8	16.9	71.8
	打牌	5	.9	.9	72.7
	散步	125	23.3	23.5	96.2
	其他	20	3.7	3.8	100.0
	Total	532	99.3	100.0	
Missing	System	4	.7		
Total		536	100.0		

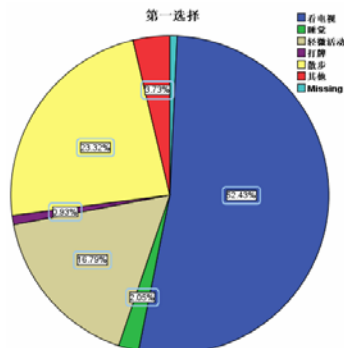


图 18-4 第一选择的饼图

表 18-6 第二选择频数分布表

第二选择		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	看电视	158	29.5	31.8	31.8
	睡觉	45	8.4	9.1	40.8
	轻微活动	70	13.1	14.1	54.9
	打牌	28	5.2	5.6	60.6
	散步	158	29.5	31.8	92.4
	其他	38	7.1	7.6	100.0
	Total	497	92.7	100.0	
Missing	System	39	7.3		
Total		536	100.0		

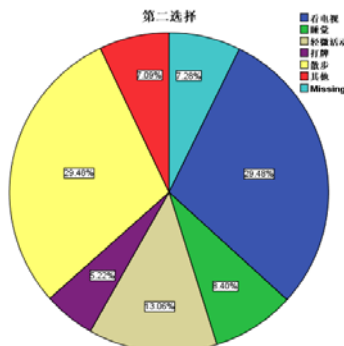


图 18-5 第二选择的饼图

表 18-7 第三选择频数分布表

第三选择		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	看电视	46	8.6	16.8	16.8
	睡觉	43	8.0	15.7	32.5
	轻微活动	68	12.7	24.8	57.3
	打牌	40	7.5	14.6	71.9
	散步	36	6.7	13.1	85.0
	其他	41	7.6	15.0	100.0
	Total	274	51.1	100.0	
Missing	System	262	48.9		
Total		536	100.0		

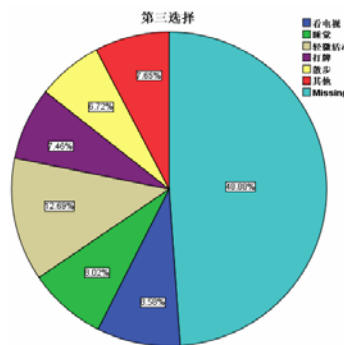


图 18-6 第三选择的饼图

(2) 组成多响应分类变量集，对该变量集进行频数分布分析

① 定义多响应分类变量集

• 按 Analyze→Multiple Response→Define Sets 顺序单击菜单项，打开对话框，将 vv1~vv3 送入 Variables in Sets 栏内。



• 在 Variables Are Coded As 栏选择 Categories，在其后输入 vv1~vv3 的取值范围，最小值 1 和最大值 6。

• 在 Name 后面输入多响应分类变量集名 rest，在下面一行输入标签“晚饭后活动”。单击 Add 将变量名送入右面的 Multi Response Sets 栏内。单击 Close 按钮。

② 进行频数分布分析

• 按 Analyze→Multiple Response→Frequencies 顺序单击菜单项，打开对话框，将多响应分类变量集\$rest 送入 Table(s) for 栏。

•单击 OK 按钮，在输出窗口中得到下面的程序和频数分布表。

③ 程序语句及语句解释

MULT RESPONSE

GROUPS=\$rest '晚饭后活动' (vv1 vv2 vv3 (1,6)) /FREQUENCIES=\$rest

MULT RESPONSE 语句调用 Mult Response 过程。

GROUPS 语句指明分组变量是已经定义的多响应分类变量集，名\$rest，变量标签是“晚饭后活动”，它由三个分类变量 vv1、vv2、vv3 组成，每个分类变量的取值范围都是 1~6。FREQUENCIES 语句是子命令，要求对等号后的多响应变量集\$rest 进行频数分布分析。

④ 运行结果见表 18-8 和表 18-9。

⑤ 结果解释

表 18-8 是观测量小结，合法观测量 533 个参与分析，占 99.4%；缺失值 3 个，被剔除占 0.6%；

表 18-9 是多响应分类变量集的频数分布表

• 左数第一列是组成多响应变量集的三个变量 vv1、vv2、vv3 的共同使用的六个值标签。

• 第三列 N 是多响应分类变量取值 1~6 即各种晚饭后活动的总频数，也就是 3 个原始变量取各代码值的总计频数。

• Total 值，答案总数是 1303。

• 第四列 Percent 中每个值是同行对应 N 值占选择答案总数 1303 的百分比。

• 第五列 Percent of Cases 中每个值是同行对应的 N 值占观测量总数的百分比。最后的百分比总和和自然会大于百分之百，因为每个回答者都有可能选择两个或三个答案。

根据第四列数据，总回答数占选择答案总数的百分比可以看出晚饭后经常看电视的人数占活动总数（答案总数）的 37.2%，散步和轻微活动的百分比分别为 24.5%和 17.5%。说明这三项活动是公务员业余生活的主要内容。轻微活动与散步总数占 41.9%比看电视的总百分比要大，说明公务员比较重视身体的活动。而其他活动（代码为 6）只占 7.6%，

表 18-8 观测量小结

Case Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
\$rest <sup>a</sup>	533	99.4%	3	.6%	536	100.0%
a. Group						

表 18-9 多响应分类变量集的频数分布表

\$rest Frequencies			
	Responses		Percent of Cases
	N	Percent	
晚饭后活动	485	37.2%	91.0%
看电视	99	7.6%	18.6%
睡觉	228	17.5%	42.8%
轻微活动	73	5.6%	13.7%
打牌	319	24.5%	59.8%
散步	99	7.6%	18.6%
其他	1303	100.0%	244.5%
Total			
a. Group			

说明活动单调。

应该说明的是该问题调查所使用的答案是经过初步调查设计的, 答案很少, 虽然排列了第一、第二、第三选择, 由于大多数人的晚饭后活动确实单调, 所以对三个变量的分析和对多响应分类变量集的频数分布分析的结果大体一致, 没有体现出多响应变量频数分布分析的特点。这里仅作为一种方法加以介绍。

## 18.4 多响应变量的交叉表分析

### 18.4.1 多响应变量集交叉表分析过程

多响应变量集交叉表分析的步骤如下:

(1) 按 **Analyze**→**Multiple Response**→**Crosstabs** 顺序单击菜单项, 打开多响应变量集的交叉表对话框, 如图 18-7 所示。在 **Multiple Response Crosstabs** 对话框中, 左上栏中显示了数据文件中所有数值型变量, 下面栏中显示了定义好的多响应变量集。

(2) 可以选择多响应变量集作为行变量送入行变量栏 **Row(s)**, 如果作为列变量则送入列变量栏 **Column(s)**, 作为层变量则送入 **Layer(s)** 栏中。

(3) 可以选择基本变量作为交叉表的行、列、层变量送入相应的变量栏中。对基本变量无论在交叉表中处于什么位置, 当选择并送入相应的栏中后需要定义它们在交叉表中出现的取值范围。

(4) 在 **Row(s)**、**Column(s)**、**Layer(s)** 栏中选择要定义分析范围的基本变量, 单击 **Define Ranges** 按钮, 打开定义范围对话框, 如图 18-8 所示。在 **Minimum** 栏中输入所选择变量的分类的最小值, 在 **Maximum** 栏中输入分类的最大值。最小值和最大值的选择可以不包括全部分类值。范围之内的分类值将出现在交叉表中。

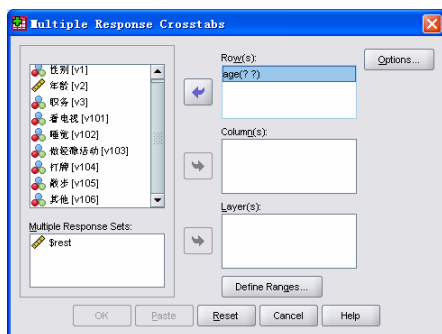


图 18-7 多响应变量集交叉表对话框

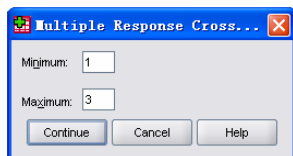


图 18-8 设置分类变量范围对话框

(5) 单击 **Options** 打开选项对话框, 如图 18-9 所示, 指定输出选项。

① 在 **Cell Percentages** 栏内选择交叉表的单元格内显示哪些统计量, 可选择输出行

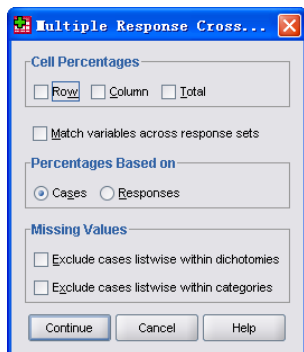


图 18-9 交叉表选项对话框

百分比 (Row)、列百分比 (Column)、总百分比 (Total)。

② **Mach variables across response sets**, 这是仅对多响应分类变量集可用的复选项。该复选项确定输出的交叉表中的对应关系。两个多响应分类变量集中的第一个集中的第一个变量与第二个集中的第一个变量作为一对, 第一个集中的第二个变量与第二个集中的第二个变量作为一对, 等等。如果选择此项, 单元格中的百分比的基数是答案总数, 而不是回答者的总数。因此该选项下面的 **Percentage Based on** 的选项中只有 **Responses** 是可以选择的, 而且是默认的。

③ 在 **Percentages Based on** 栏中选择:

- **Cases**, 交叉表中各单元格中的百分比的计算基数是观测量数, 即答卷人数。
- **Responses**, 交叉表中各单元格中的百分比的计算基数是总答案数。由于是多项选择题, 因此一般情况是答案数大于回答者人数。

④ **Missing Values** 栏 选择项的含义与多响应变量集的频数分布的一致。参见本章第 18.3.1 节的相关内容。

## 18.4.2 多响应二分变量集的交叉表分析实例

【例 3】仍以数据文件 data18-01 数据为例加以说明。想分析 50 岁以下的各年龄段, 饭后做什么。在数据文件中, 原来只记录了年龄, 为了分析, 先根据年龄变量 **V02**, 生成年龄段变量 **age**。利用 **Transform** 菜单中的 **Compute** 功能, 按下列原则将年龄变量分段:

if  $V02 \leq 17$  then  $age=0$ ; 不参与分析没有定义值标签, 可以当作缺失值处理。

if  $17 < V02 \leq 30$  then  $age=1$  值标签为 “ $\leq 30$ ”; 以下叙述中称之为 30 岁以下。

if  $30 < V02 \leq 40$  then  $age=2$ ; 值标签为 “31~40”

if  $40 < V02 \leq 50$  then  $age=3$ ; 值标签为 “41~50”

if  $V02 > 50$  then  $age=4$ ; 值标签定义为 “>50”

1. 使用多响应变量分析功能进行交叉表分析

(1) 进行交叉表分析的步骤如下:

先定义多响应二分变量集 **\$rest**。方法见 18.2 节。然后进行如下操作:

① 按 **Analyze**→**Multiple Response**→**Crosstabs** 顺序单击菜单项, 打开多响应变量集的交叉表对话框。

② 要做二维交叉表, 选择年龄段变量 **age** 送入行变量栏 **Row(s)**;

③ 单击 **Define Ranges** 按钮, 输入最小值为 1, 最大值为 3。

④ 选择多响应二分变量 **\$rest** 送入 **Column(s)** 栏。

⑤ 单击 **Options** 按钮, 选择 **Row(s)**、**Column(s)**、**Total**, 要求每个单元格除显示单

元格频数外，都显示行百分比、列百分比和总百分比。

由于多响应变量集是由若干二分变量组成的，因此不能选择 **Mach** 项。计算百分比的基数是观测数。选择 **Percentage Based on** 栏中的 **Cases**。

(2) 交叉表分析的程序语句如下：

```
MULT RESPONSE                                ①
GROUPS=$rest '饭后活动' (V101 V102 V103 V104 V105 V106 (1)) ②
/VARIABLES=age(1 3)                            ③
/TABLES= age BY $rest                          ④
/CELLS=ROW COLUMN TOTAL                       ⑤
/BASE=cases .                                  ⑥
```

语句说明

① 调用 **Mult Response** 过程进行多响应变量集的分析。

② 多响应变量集 **\$rest**，标签为“饭后运动”，由 **V101～V106** 六个基本变量组成。在输出表中表示的频数是这些变量值为 1 的计数值。

③ 基本变量为 **age**，其分析的值范围是 1～3。

④ 要求输出交叉表，**BY** 前是行变量，行变量是基本变量 **age**，列变量在 **BY** 后面，是多响应二分变量集 **rest**。

⑤ 在交叉表中要求输出每个单元格频数的行百分比、列百分比和总百分比。

⑥ 所有百分比值的基数均为观测量总数。如果以答案总数为基数，语句为 **/BASE=RESPONSES**。

(3) 输出结果见表 18-10～表 18-12。

(4) 输出解释

表 18-10 为观测量小结，注意凡是没有包括在分析范围之内的观测量，例如小于 17 岁和大于 50 岁的都计数在缺失值内。因此合法值为 468。

表 18-11 是以观测量总数为百分比基数的交叉。表中每个单元格中的数据自上至下为：该单元格的频数、行百分比、列百分比、总百分比。以左上角第一个单元格为例给出解释：该单元格表示 30 岁以下的人晚饭后看电视的 59 人，占这个年龄段 93 人的 63.4%，占晚饭后看电视总人数（三个年龄段）333 人的 17.7%，占三个年龄段总人数 468 的 12.6%。

最右边一列的各单元格中，上边的数值是该行观测量总数，第二个数值是该行观测量总数占总观测量数的百分比。例如年龄大于等于 18 岁，小于等于 30 岁的 93 人，注意因为允许多项选择，这个 94 并不等于该行中各单元格计数 **Count** 之和。这个年龄段的 93 人占总人数 468 人的 19.9%。

表下面第一行 **Count** 为各列观测量总数，就是该列各单元格中计数 **Count** 的总和。% of

表 18-10 观测量小结表

Case Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
age*\$rest	468	87.3%	68	12.7%	536	100.0%

Total是列观测量数占总观测量数的百分比。例如，晚饭后看电视的333人占观测量总数468的71.2%。

右下角单元格中显示的是观测量总数468，并表示以该数值为分母计算的各百分比。  
注意，交叉表中的列变量是多响应二分变量集，因此最后右下角单元格中的总计数468不等于总最后一行上各列答案计数Count的总和704，而是总观测量数468；最后一行上的各列百分比% of Total的总和自然也不是100%，而是总答案数除以总观测量数的商，大于100%，即 $704/468=150\%$ ，这个数在这里未列出。

比较表18-11和表18-12可以看出各对应单元格中的频数是一样的，但是对应单元格中的行、列和总百分比是不同的。

表 18-11 以观测量总数为百分比基数的交叉表

		age*strest Crosstabulation						Total
		饭后活动*						
		看电视	睡觉	做轻微活动	打牌	散步	其他	
年龄段 <=30	Count	59	8	26	5	31	10	93
	% within age	63.4%	8.6%	28.0%	5.4%	33.3%	10.8%	
	% within \$strest	17.7%	16.7%	22.0%	29.4%	21.2%	23.8%	
	% of Total	12.6%	1.7%	5.6%	1.1%	6.6%	2.1%	19.9%
31-40	Count	127	19	42	5	51	21	180
	% within age	70.6%	10.6%	23.3%	2.8%	28.3%	11.7%	
	% within \$strest	38.1%	39.6%	35.6%	29.4%	34.9%	50.0%	
	% of Total	27.1%	4.1%	9.0%	1.1%	10.9%	4.5%	38.5%
41-50	Count	147	21	50	7	64	11	195
	% within age	75.4%	10.8%	25.6%	3.6%	32.8%	5.6%	
	% within \$strest	44.1%	43.8%	42.4%	41.2%	43.8%	26.2%	
	% of Total	31.4%	4.5%	10.7%	1.5%	13.7%	2.4%	41.7%
Total	Count	333	48	118	17	146	42	468
	% of Total	71.2%	10.3%	25.2%	3.6%	31.2%	9.0%	100.0%

Percentages and totals are based on respondents.  
a. Dichotomy group tabulated at value 1.

表 18-12 以答案总数为百分比基数的交叉表

		age*strest Crosstabulation						
		饭后活动*						
		看电视	睡觉	做轻微活动	打牌	散步	其他	Total
年龄段 <=30	Count	59	8	26	5	31	10	139
	% within age	42.4%	5.8%	18.7%	3.6%	22.3%	7.2%	
	% within \$strest	17.7%	16.7%	22.0%	29.4%	21.2%	23.8%	
	% of Total	8.4%	1.1%	3.7%	.7%	4.4%	1.4%	19.7%
31-40	Count	127	19	42	5	51	21	265
	% within age	47.9%	7.2%	15.8%	1.9%	19.2%	7.9%	
	% within \$strest	38.1%	39.6%	35.6%	29.4%	34.9%	50.0%	
	% of Total	18.0%	2.7%	6.0%	.7%	7.2%	3.0%	37.6%
41-50	Count	147	21	50	7	64	11	300
	% within age	49.0%	7.0%	16.7%	2.3%	21.3%	3.7%	
	% within \$strest	44.1%	43.8%	42.4%	41.2%	43.8%	26.2%	
	% of Total	20.9%	3.0%	7.1%	1.0%	9.1%	1.6%	42.6%
Total	Count	333	48	118	17	146	42	704
	% of Total	47.3%	6.8%	16.8%	2.4%	20.7%	6.0%	100.0%

Percentages and totals are based on responses.  
a. Dichotomy group tabulated at value 1.

表18-12是以答案总数为计算百分比基数的交叉。以左上角单元格为例，小于30岁大于等于18岁的人，晚饭后看电视人数为59人，占这个年龄段回答者总人数（行总和）139人的42.4%（行百分比）；该频数占选择看电视总人数(列总和)333人的17.7%（列百分比）；占总答案数704的8.4%。该频数占回答所在列表示的活动类型的人数的百分比和占总选择数的百分比。最右边一列中各单元格上边数值是各行频数之和，下边的数值是各行频数之和占总答案数704的百分比。第一行总频数，即总选择数139，占704的19.7%。

表最下面一行数值是各种选择（各列）的总频数，这与表18-11是相同的，而百分比却是各列总频数占总选择数704的百分比。

总答案数在右下角中是704，它是各列总频数Count数值之和。列百分比之和是100%，同时也是行百分比之和。

读者可以根据对各项的上述解释，观察交叉表，得出必要的结论。  
多响应分类变量集的交叉表分析的操作方法与多响应二分变量集的交叉表分析方法相同。读者可以自己实践，此处不再赘述。

注意：使用多响应变量集交叉表分析功能也可以做单个基本变量间的交叉表。但其

功能不如Descriptive Statistics 二级菜单中的Crosstabs项的交叉表分析功能强。如果希望进行卡方检验,或得到表明分布情况的图形,分析原变量的频数分布和得到交叉表,还是应该使用Descriptive Statistics 二级菜单中的Crosstabs项的交叉表分析功能。

## 18.5 多响应变量集分析的过程语句

在窗口方式中可选择的功能是有限的,许多功能还需要在主对话框中,通过Paste按钮生成程序,再根据各种命令语句及其选择项将源程序完善,以便得到更满意的输出结果。

### 1. 过程语句格式

#### MULT RESPONSE

```

{/GROUPS=groupname['label'](varlist({value1,value2})){value}... [groupname...]
  {/VARIABLES=varlist(min,max) [varlist... ]
  {/FREQUENCIES=varlist }
  {/TABLES=varlist BY varlist...[BY varlist][(PAIRED)]}{/varlist BY...}
  [/MISSING={TABLE**}{MDGROUP}{MRGROUP}] [INCLUDE]
  [/FORMAT={LABELS**}{NOLABELS}{CONDENSE}{ONEPAGE}{TABLE**}[DOUBLE] ]
  [/BASE={CASES** }{RESPONSES}]
  [/CELLS={COUNT**} [ROW] [COLUMN] [TOTAL] [ALL]]

```

标有“\*\*”的参数是当所在子命令在程序中不出现时,执行该子命令使用的默认值。

### 2. 基本要求

(1) MULT RESPONSE命令语句调用MULT RESPONSE过程进行多响应变量集的分析。此语句是必须有的,而且必须在程序第一行出现。

(2) GROUPS和VARIABLES两个子命令必须在程序中至少出现一个。GROUPS子命令定义要进行分析的多响应变量集,VARIABLES子命令指定参与分析的单一分类变量或二分变量。无论使用了GROUPS子命令还是VARIABLES子命令,定义的参与分析的变量或多响应变量集必须在两个以上。

(3) FREQUENCIES、TABLES两个子命令必须在程序中至少出现一个。FREQUENCIES子命令要求对GROUPS指定的多响应变量集或VARIABLES子命令指定的单一分类变量进行频数分布分析,输出频数分布表。TABLES子命令则要求对GROUPS或VARIABLES子命令指定的多响应变量集或单一变量进行交叉表分析,输出交叉表。

(4) 语句的顺序要求:其他子命令使用时只能出现在上述4个子命令后面。

### 3. 子命令

(1) GROUPS子命令,定义多响应二分变量集和多响应分类变量集

① 定义多响应二分变量集的格式是:

**GROUPS=**多响应二分变量集名[标签](组成变量集的二分变量列表)(二分变量中要计数的值)

除标签是可选外,其他部分是必需的。组成二分变量集的二分变量列表、二分变量中要计数的值(一般采用表示“是”的值),是必须指定的子命令参数,要放在小括号中。

② 定义多响应分类变量集的格式是:

**/GROUPS=**多响应分类变量集名[标签](组成变量集的分类变量列表)(分类变量的取值列表)

其中标签是可选的,其他部分是必需的。分类变量取值必须一个个列出。如果是有序的,应该按顺序列出。变量列表和取值列表分别使用小括号。多响应变量集名字母部分前加符号\$。该子命令中指定的变量集数不能超过20个,该子命令中指定的变量数与**VARIABLES**命令中的变量数加起来不能超过100个,其中的分类变量的取值个数不能超过32767个,必须是整数,否则系统自动截尾。

### (2) **VARIABLES**子命令

该子命令定义参与分析的单一变量(原变量)格式是:

**VARIABLES=**变量列表(最小值,最大值)[变量列表(最小值,最大值)...

可以出现若干个变量列表,相同取值范围的变量列在一起,只在后面出现一次取值范围即可。指定的变量数与**GROUPS**子命令指定的变量数之和至多不能超过100个,变量分类不能超过32767个。

### (3) **FREQUENCIES**子命令

格式是:**FREQUENCIES=**变量或多响应变量集列表。

该子命令对出现在等号后面的每个变量或多响应变量集分别做频数分布分析。等号后面的多响应变量集必须要在该子命令前使用**GROUPS**子命令定义过。

### (4) **TABLES**子命令

① 格式是:**TABLES=**行变量 **BY** 列变量 [**BY** 层变量...][**(PAIRED)**]

② 该子命令要求输出交叉表,并定义交叉表的行变量、列变量、层变量……最多定义五维交叉表,最多定义10个交叉表。

③ 系统限制,在**TABLES**和**FREQUENCIES**两个子命令中输出的一个表中最多有200个非空行和200个非空列。两个命令中出现的多响应变量和单一变量总数不能超过100个。

④ 可选项还有关键字**PAIRED**。

• 什么情况使用**PAIRED**。对两个多响应变量集输出交叉表时,默认的输出是把第一个变量集中的每个变量与第二个变量集中的每个变量作成交叉表,而且对每个单元格进行计总和频数。这样,某些答案在表中可能出现多于一次。使用**PAIRED**把第一个多响应变量集中的第一个变量与第二个多响应变量集中的第一个变量配对,第一个多响应变量集中的第二个变量与第二个多响应变量集中的第二个变量配对,以此类推。

- 关键字 **PAIRED** 位置。在 **TABLES** 子命令中，在一个指定的表中最后一个变量名后面，放在括号里。

- 在输出中的位置。当要求交叉表配对时，多响应变量集成员变量按 **GROUPS** 子命令中出现的顺序出现在交叉表中。虽然在包括单一变量和多响应二分变量集的表中可以要求配对，但只有在多响应分类变量集中的变量才配对。输出中的配对表标有 **PAIRED GROUP** 标签。

- 在交叉表中的百分比总是以答案数为基数，而不是以观测量数为基数。

#### (5) CELL子命令

该子命令是可选的，格式是：

**[/CELLS=[COUNT\*\*] [ROW] [COLUMN] [TOTAL] [ALL]]**

① 如果不使用该子命令，交叉表的单元格中只显示该单元格的频数。

② 使用该子命令，子命令可以带有以下参数，均为可选项。

- **COUNT** 系统默认的选项，要求输出交叉表的单元格中显示该单元格的频数。只要求使用 **CELL** 子命令，此项不可以省略。
- **ROW** 要求输出的交叉表单元格中显示行百分比。
- **COLUMN** 要求输出的交叉表单元格中显示列百分比。
- **TOTAL** 要求输出的交叉表单元格中显示总百分比。
- **ALL** 要求输出的交叉表单元格中显示：单元格频数、行百分比、列百分比、总百分比。

#### (6) BASE子命令

① 子命令格式：**[/BASE={CASES\*\*}]{RESPONSES}]**

② 该子命令指定在计算单元格百分比和边际百分比时的分母，是可选子命令。

③ 不使用该子命令，并且不要求做配对交叉表时，计算百分比时使用观测量数做分母，与使用该子命令 **BASE=CASES** 时是相同的。如果在 **TABLES** 子命令中指定了 **PAIRED** 关键字，则在 **BASE** 子命令中不能指定该选项。

④ **BASE=RESPONSES** 计算百分比时使用答案数做分母。要求做配对交叉表时，该选项是系统默认的计算方法。

#### (7) MISSING子命令

① 子命令格式：**[/MISSING={TABLE\*\*}{MDGROUP}{MRGROUP}] [INCLUDE]**

② 是可选子命令。允许在程序中只出现命令关键字。子命令中每个参数都是可选项。

③ 系统默认处理带缺失值观测量的方法是：带有指定范围外的变量值观测量不出现在输出表中，但包括在缺失类中。这样指定一个把缺失值剔除在外的范围等价于默认的对缺失值的处理。

④ 系统默认：对多响应二分变量集，其组成变量中没有一个包含指定的计数值的观测量，将从输出表的计数中剔除。对多响应分类变量集，如果一个观测量的多响应分类



变量集中的各变量合法值没有一个在指定的范围内，默认该观测量是缺失的，并剔除。

⑤ MISSING子命令中可以选用的参数：

- **TABLE** 逐个表格确定缺失值，根据表格成分剔除缺失值。如果不使用 MISSING 子命令，这是默认的。
- **MDGROUP** 多响应二分变量集中的组成变量带有缺失值的观测量从多响应二分变量集分析表的单元格计数中剔除。
- **MRGROUP** 多响应分类变量集中的组成变量带有缺失值的观测量从频数表或交叉表分析的单元格计数中剔除。
- **INCLUDE** 读者缺失值如果包括在 **GROUPS** 或 **VARIABLES** 子命令中指定范围内，做合法值处理。

(8) **FORMAT**子命令

① 子命令格式：**[/FORMAT={LABELS\*\*}{NOLABELS}{CONDENSE}{ONEPAGE }  
{TABLE\*\* } [DOUBLE]]**

② **FORMAT** 控制表的输出格式。单个关键字也可以作为一个子命令出现。标签由两个关键字控制。

- **LABELS**，系统默认。要求在输出的频数表或交叉表中显示值标签。
- **NOLABELS**，要求在频数分布表或交叉表中不显示多响应分类变量和单一变量的值标签。不能压缩掉多响应二分变量集中用做值标签的变量标签。

③ 下列关键字可以用于频数表的格式中。

- **DOUBLE**，系统默认 **MULT RESPONSE** 使用单倍间隔。此选项使频数表中使用双倍间隔。
- **TABLE**，如果不使用 **FORMAT** 子命令，使用一列压缩格式的频数表。这是默认的格式。
- **CONDENSE**，对频数表使用压缩格式。该选项对所有多响应分类变量集和单一变量的频数表使用 3 列压缩格式，不显示标签。该选择项不适用于多响应二分变量集。
- **ONEPAGE**，有条件地使用频数表的压缩格式。如果在一页不能输出整个表格，使用 3 列压缩格式。该选项不适用于多响应二分变量集的分析。

## 18.6 使用Table功能分析多响应变量集

使用 **Analyze** 的 **Table** 功能也可以分析多响应变量集，得到很好形式的表格。要想使用 **Tables** 菜单中的 **Custom Tables** 功能做表并进行分析。

### 18.6.1 简单频数分布分析

使用 Analyze→Tables→Custom tables 功能做表, 进行频数分布分析。

(1) 多响应二分变量集的分析, 以 data18-01 为例。

① 打开 Custom Tables 对话框, 见图 18-10, 把多响应变量集拖曳到 Columns (或 Rows) 栏中。

② 单击 Titles 选项卡, 在 Titles 栏中输入表格标题“某单位公务员晚饭后活动类型频数分布表 ( )”。鼠标光标置于括号中, 单击 Date 图标, 要求显示分析日期。

③ 在主对话框左下角 define 栏中, 单击 N% Summary Statistics 按钮, 在相应对话框, 如图 18-11 中的 Statistics 栏中选择 Row N % 送入右边的 Display 栏中。注意这是以总观测量数为基数计算百分比的。单击 Apply selections 按钮, 回到主对话框的 Table 选项卡。

④ 单击左下角的 Categories and Total 按钮, 打开相应对话框, 见图 18-12。

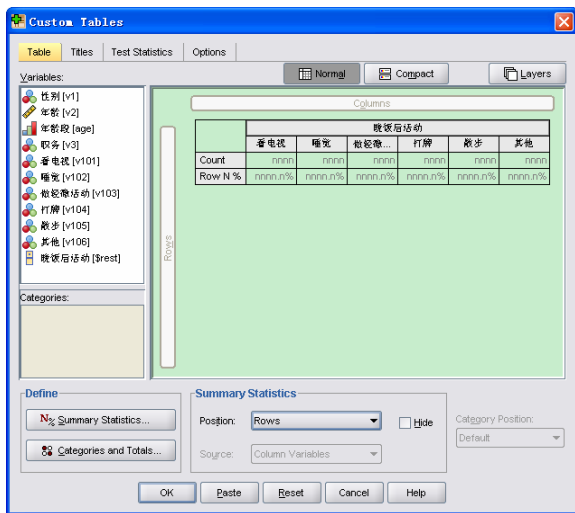


图 18-10 Custom tables 功能主对话框

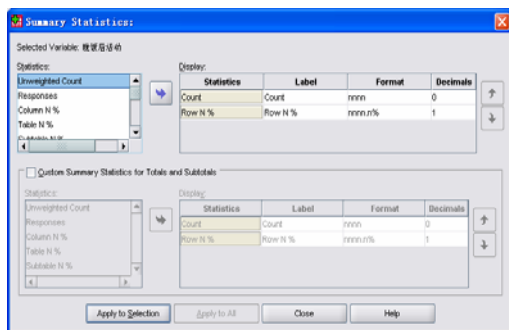


图 18-11 综合统计量对话框

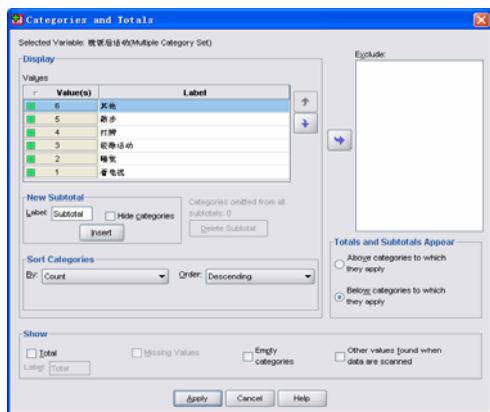


图 18-12 分类汇总对话框

在 Sort Categories 下的 By 选项框的下拉列表中选择 Count, 并在 Order 下拉列表中选择 Descending 要求按单元中的频数降序排序。但注意 SPSS 版本, 如果在选择项下面

有一行说明：“The grid may not reflect the order in which categories will appear in the table”。则这个功能可能不能实现。

将 Show 栏的各项前边的对钩去掉，要求不显示空单元。也不选择 Total，因为对于多响应问题这里也显示总观测量数，而不是各项频数的总和，而各项频数的总和也没有什么实际意义故不选。单击 Apply 按钮，返回主对话框。

⑤ 在主对话框中的 Summary Statistic 栏中选择 Rows，要求把统计量标题排在表格左边，每类统计量占一行。结果见表 18-13。

(2) 多响应分类变量集与二分变量集的操作步骤完全相同，数据文件 data18-02。得到的表格是按前三种选择的频数分布，见表 18-14。

表 18-13 二分多响应集的频数分布分析结果

表 18-14 多响应分类变量集的频数分布分析结果

某单位公务员晚饭后活动类型频数分布表（2009-5-2）						
	晚饭后活动					
	看电视	散步	做轻微活动	睡觉	其他	打牌
Count	381	175	127	57	44	18
Row N %	71.8%	33.0%	23.9%	10.7%	8.3%	3.4%

某单位公务员晚饭后活动类型频数分布表（2009-5-2）						
	晚饭后活动					
	看电视	睡觉	轻微活动	打牌	散步	其他
Count	485	99	228	73	318	99
Row N %	91.0%	18.6%	42.8%	13.7%	59.7%	18.6%

18.6.2 交叉表分析

【例 4】仍以 data18-01、data18-02 数据为例使用 Analyze→Tables 中的 Custom Tables 功能分析多响应变量的交叉表。

1. 多响应二分变量集的交叉表

(1) 打开数据文件 data18-01，按 Analyze→Tables→Multiple Response Sets 打开定义多响应变量集的对话框，定义二分变量集\$rest。具体方法见 18.2 节。使用这个菜单定义的多响应变量集是可以保存的，到下一次打开 SPSS 时仍然存在，不用重新定义。

(2) 按 Analyze→Tables→Custom Tables 顺序打开如图 18-10 的对话框，将\$rest 拖曳到 Columns 栏中，进行以下定义：

① 在主对话框左下角，单击 N% Summary Statistics 按钮，在相应对话框（见图 18-11）的 Statistics 栏中选择以下各常用项，送入 Display 栏中。

- Count 要求显示每个单元格的计数，即每个年龄段，晚饭后从事各种活动的人数。
- Row N % 行百分比。是 Count 做分母的百分比。
- Response 回答人数。从数值上与 Count 相等，但总计、百分比有区别。
- Row response % 各单元格中的回答人数相对于该行（同年龄段）总回答人数的百分比。
- Table N % 该单元格的中的 count 频数，与样本量 Count 的百分比值。

单击 Apply Selections 按钮，回到主对话框的 Table 选项卡。

② 单击 Categories and Totals，在相应的对话框中（见图 18-12）的 Sort Categories By 栏选择 Count，Order 栏选择 Descending。要求生成的表格按单元格频数降序排列；在

Show 栏选择 Total 要求显示总计，去除 Empty Categories 复选项，要求不显示空单元。单击 Apply，将设置施加到 \$rest 变量上。返回主对话框 Table 选项卡。

(3) 将 age 年龄段变量拖曳到 Rows 行上

① 单击 Categories and Totals，进入相应的对话框，在 Display 栏中选择 Value 值为 3（标签 41~50）组，在 New Subtotal 栏中单击 Insert 按钮。要求在该组后面加一个阶段总计。因为，该研究更关心 50 岁以下的公务员饭后活动情况。在 Show 栏中只选择 Total 复选项，要求显示总计。单击 Apply 将以上设置施加于行变量上，返回主对话框 Table 选项卡。

② 在主对话框中的 Summary Statistics 栏设置 Position 设置为 Rows，Source 设置为 Column Variable；即要求将列变量的综合统计量显示在行上。

(4) 程序语句

\* Custom Tables. CTABLES /VLABELS VARIABLES=\$rest age DISPLAY=DEFAULT

/TABLE age BY \$rest [COUNT F40.0, ROWPCT.COUNT PCT40.1, RESPONSES F40.0,

ROWPCT.RESPONSES PCT40.1, TABLEPCT.COUNT PCT40.1]

/SLABELS POSITION=ROW /CATEGORIES VARIABLES=\$rest ORDER=D KEY=COUNT

EMPTY=EXCLUDE TOTAL=YES POSITION=AFTER

/CATEGORIES VARIABLES=age [1, 2, 3, SUBTOTAL, 4] EMPTY=EXCLUDE

TOTAL=YES POSITION=AFTER

(5) 运行该程序，得到结果见表 18-15。

表 18-15 用 Table 功能实现交叉表

		晚饭后活动						Total
		看电视	散步	做轻微活动	睡觉	其他	打牌	
年龄段 <=30	Count	63	33	27	9	10	5	97
	Row N %	64.9%	34.0%	27.8%	9.3%	10.3%	5.2%	100.0%
	Responses	63	33	27	9	10	5	147
	Row Responses %	42.9%	22.4%	18.4%	6.1%	6.8%	3.4%	100.0%
31~40	Table N %	11.9%	6.2%	5.1%	1.7%	1.9%	.9%	18.3%
	Count	127	51	42	19	21	5	180
	Row N %	70.6%	28.3%	23.3%	10.6%	11.7%	2.8%	100.0%
	Responses	127	51	42	19	21	5	265
41~50	Row Responses %	47.9%	19.2%	15.8%	7.2%	7.9%	1.9%	100.0%
	Table N %	23.9%	9.6%	7.9%	3.6%	4.0%	.9%	33.9%
	Count	147	64	50	21	11	7	195
	Row N %	75.4%	32.8%	25.6%	10.8%	5.6%	3.6%	100.0%
Subtotal	Responses	147	64	50	21	11	7	300
	Row Responses %	49.0%	21.3%	16.7%	7.0%	3.7%	2.3%	100.0%
	Table N %	27.7%	12.1%	9.4%	4.0%	2.1%	1.3%	36.7%
	Count	337	148	119	49	42	17	472
>50	Row N %	71.4%	31.4%	25.2%	10.4%	8.9%	3.6%	100.0%
	Responses	337	148	119	49	42	17	712
	Row Responses %	47.3%	20.8%	16.7%	6.9%	5.9%	2.4%	100.0%
	Table N %	63.5%	27.9%	22.4%	9.2%	7.9%	3.2%	88.9%
Total	Count	381	174	127	57	44	18	531
	Row N %	71.8%	32.8%	23.9%	10.7%	8.3%	3.4%	100.0%
	Responses	381	174	127	57	44	18	801
	Row Responses %	47.6%	21.7%	15.9%	7.1%	5.5%	2.2%	100.0%
Table N %		71.8%	32.8%	23.9%	10.7%	8.3%	3.4%	100.0%

可以看出,使用table功能做多响应变量的交叉表可以插入其他统计量,例如分段的小结Subtotal。它是对所在行以上的小结。即50岁以下公务员的饭后活动的小结。选择看电视的人337人,占50岁以下回答者总人数472人的71.4%。选择轻微活动和散步的人总共267人占50岁以下回答总人数472的56.6%。相比之下做轻微活动和散步的人比例不如看电视的大。

## 2. 多响应分类变量集的交叉表

【例5】以数据data18-02为例做多响应分类变量集的交叉表。

(1) 打开数据文件data18-02,按Analyze→tables→Multiple Response Sets打开定义多响应变量集的对话框,定义多响应变量集rest。具体方法见18.2节。

(2) 按Analyze→Tables→Custom Tables顺序打开如图18-10的对话框,将rest拖曳到Columns栏中。单击Categories and Total,在相应的对话框中(见图18-12)的Sort Categories By栏选择Cell Count,Order栏选择Descending。要求生成的表格按单元格降序排列;在Show栏选择Total要求显示总计。单击Apply,将设置施加到rest行变量上。返回主对话框Table选项卡。

(3) 将age年龄段变量拖曳到Rows行上

① 单击Categories and Totals,进入相应的对话框,在Display栏中选择Value值为3(标签41~50)组,在New Subtotal栏中单击Insert按钮。要求在该组后面加一个阶段总计。因为,该研究更关心50岁以下的公务员饭后活动情况。在Show栏中只选择Total复选项,要求显示总计。单击Apply将以上设置施加于行变量上,返回主对话框Table选项卡。

② 在主对话框中的Summary Statistics栏的Position设置为Rows,Source设置为Column Variable;即要求将列变量的综合统计量显示在行上。

(4) 将性别变量拖曳到Layer图标上。要求按性别分页做交叉表。单击Categories and Total进入相应的对话框,选择Show中的Total,其他不选择。单击Apply返回主对话框。

单击Past在语句窗口生成如下程序。读者可以参照第6章有关内容理解该程序,在此不再解释。

## (5) 程序语句

CTABLES

```
/VLABELS VARIABLES=$rest age V1 DISPLAY=DEFAULT
```

```
/TABLE age [COUNT F40.0] BY $rest BY V1
```

```
/SLABELS POSITION=ROW
```

```
/CATEGORIES VARIABLES=$rest [2, 3, 4, 5, 6, 1] EMPTY=EXCLUDE TOTAL=YES
```

```
POSITION=AFTER
```

```
/CATEGORIES VARIABLES=age [1, 2, 3, SUBTOTAL, 4] EMPTY=EXCLUDE
```

```
TOTAL=YES POSITION=AFTER
```

```
/CATEGORIES VARIABLES=V1 [1, 2] EMPTY=EXCLUDE.
```

(6) 运行程序生成表 18-16 和表 18-17 的输出结果。

比较Table菜单的Custom Tables功能的表格和Multiple Response 多响应变量分析中的表格,可以看出Table的制表功能更强,但是无论是Table菜单的Custom Tables还是Multiple Response多响应变量分析,对多响应变量集,无论是二分集还是分类集,都不能做任何检验。只能做各种百分比分析。

表 18-16 第一页交叉表

性别 男			晚饭后活动					
			睡觉	轻微活动	打牌	散步	其他	看电视
年龄 段	<=30	Count	16	24	12	19	12	43
	31~40	Count	15	33	12	47	21	82
	41~50	Count	8	46	15	76	15	97
	Subtotal	Count	39	103	39	142	48	222
	>50	Count	4	17	4	33	6	37
	Total	Count	43	120	43	175	54	259
								Total
								51
								89
								102
								242
								40
								282

表 18-17 第二页交叉表

性别 女			晚饭后活动					
			睡觉	轻微活动	打牌	散步	其他	看电视
年龄 段	<=30	Count	20	18	5	21	1	42
	31~40	Count	20	36	7	49	23	79
	41~50	Count	12	45	16	59	18	87
	Subtotal	Count	52	99	28	129	42	208
	>50	Count	4	9	2	14	3	18
	Total	Count	56	108	30	143	45	226
								Total
								46
								91
								94
								231
								20
								251

## 习 题 18

1. 多响应变量分几种类型, 各对应哪种分析方法?
2. 在分析多响应变量之前, 对原始数据应该进行怎样的处理?
3. 设计两种问卷对中央电视台的 10 个频道的认知情况进行调查, 用两种方式建立数据文件, 并对其知名度进行排序。

提示: 问卷 1: 请说出你所知道的中央电视台的频道名称: 顺序记录频道号和名称

问卷 2: CCTV1 的频道名称你知道吗? 答对记 1, 不对记 0

CCTV2 的频道名称你知道吗? 答对记 1, 不对记 0

CCTV3

.....

# 第 19 章 生存分析

## 19.1 生存分析概述

### 19.1.1 生存分析与生存数据

生存分析广泛应用于生物医学、工业、社会科学、商业等领域，例如肿瘤患者经过治疗后生存的时间、电子设备的寿命、罪犯假释的时间，婚姻的持续时间、保险人的索赔等。这类问题数据特点是在研究期间结束时，所要研究的事件还没有发生，或过早终止，使要收集的数据发生缺失。这样的数据称为生存数据。生存分析就是要处理、分析生存数据。

#### 1. 生存分析的类型与 SPSS 过程

生存分析方法可分为三类：非参数法、生命表法和乘积极限法（用于估计生存率），Log-rank 检验用于单因素预后分析；半参数模型即比例风险模型，辨别多协变量的预后因素；参数法一般也用作预后分析，如指数模型，Weibull 模型等。如果了解生存数据服从某特定分布，那么参数检验比非参数和半参数检验更有效。

在 Analyze 菜单下的 Survival 子菜单中，提供了 Life table、Kaplan-Meier、Cox Regression、Cox w/Time-Dep Cov（带时间相依性变量的生存分析）4 种生存分析方法。

#### 2. 生存分析的数据

生存数据包括生存时间以及与其相关因素。生存数据有一个最重要的特点：在研究期间结束时在某些个体身上还没有发生。所观测的含有这些事件的数据称为删失数据（Censored Data）。如果生存数据中没有删失观测，该生存数据称为完全数据。

在删失数据中按照删失数据发生的时间，可分为右删失、中间删失和左删失。右删失类型包括单式删失和随机删失，常用的删失类型主要有 I 型和 II 型删失，它们均属于单式删失。动物实验、设备寿命研究中常遇到删失数据。

由于时间和费用受到限制，研究者常常不能等到所有动物死亡。事先确定截止观测的日期称为 I 型删失，又称定时删失。如果选择试验进行到有一固定数目的动物死亡为止，称为 II 型删失，又称定数删失。一般在数据的右上角标注“+”号表示是删失数据。

### 19.1.2 生存时间函数

生存时间测量某事件出现的时间。通常用下列 3 个函数来描述，生存函数、概率密

度函数和危险率函数。它们在数学上是等价的, 得出其中一个, 可以推导出另两个。

生存函数 (Survival function), 又称累积生存率, 记作  $S(t)$ , 它是指个体生存时间长于  $t$  的概率, 即

$$S(t) = P(\text{个体生存时间长于 } t)$$

概率密度函数 (Probability Density Function), 又称密度函数, 记作  $f(t)$ ,  $f(t)$  的图形叫做密度曲线, 在任何时间区间内死亡的比例和死亡出现机会的峰值都可以从密度曲线找出, 函数表达式为

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(\text{个体在区间}(t, t+\Delta t)\text{中死亡})}{\Delta t}$$

危险率函数 (Hazard Function), 又称为风险函数、瞬间死亡率、死亡强度、条件死亡率、分年龄死亡率、危险率, 记为  $h(t)$ , 危险率函数是生存分析最基本的函数, 即

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(\text{年龄是 } t \text{ 的个体在}(t, t+\Delta t)\text{中死亡})$$

### 19.1.3 Cox回归模型

当众多的危险因素对生存时间有影响时, 应关心这其中哪些危险因素对生存时间有重要的影响, 也就是确认重要的预后因素。通过建立生存时间随危险因素变化的回归模型, 来确定这些对生存时间有影响的预后因素, 并根据危险因素在模型中的影响对生存率进行预测。但是危险率往往难以估计, 所以不宜采用非参数或参数模型方法。1972年英国统计学家 D.R.Cox 提出了比例风险模型 (the Proportional Hazard Model, PHREG), 该模型可以很好地解决上述问题, 故又称为 Cox 回归或 Cox 模型。Cox 模型在表达形式上与参数模型相似, 但对各参数进行估计时又不依赖特定分布的假设, 所以也称为半参数回归模型。当生存时间是连续分布且预后变量间相互作用可被忽视时危险率  $h(t)$  为

$$h(t) = h_0(t)e^{(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}$$

在方程式中  $h_0$  是基准的生存分布的危险率函数,  $\beta$  是回归系数,  $x$  为预后变量。由于 Cox 模型假设, 每个预后变量的危险率在时间上正比于基础危险率 ( $h_0$ ), 从而无需计算 ( $h_0$ )。

Cox 模型除辨认预后因素外, 还可以确定预后指数或比率, 即求每个个体的  $\log_e[hi(t)/h_0(t)]$ 。

## 19.2 生命表分析

### 19.2.1 生命表分析概述

生命表 (Life Table, LT) 又称寿命表, Mantel 和 Haznszel (1959 年) 提出用生命表的方法可以比较两种生存模式。生命表分析方法是用来测定死亡率和描述群体生存现



象的。一般说来，生命表用来概括在特定的时期里特定人口的死亡情况，统称为人口生命表。生命表应用于患有某种疾病并且在一定时期受到跟踪研究的患者身上，对患者构造出的生命表为临床生命表。人口生命表和临床生命表在计算方法上是相似的，但所要求的数据的来源不同。

生命表用于大样本，并且对生存时间的分布不限，这也是它的优点，所以它是目前广泛应用的一种非参数分析方法。在生命表中，生存函数和生存率的估计依赖于生命表中所有的区间。如果每个区间都很短，则区间个数很多，计算工作变得很繁重，不能体现其优点。尽管利用计算机分析使这项工作轻松简单，但输出的结果却十分冗长。用于生命表的一个假定是总体在每个区间内各处有近似相等的生存概率。如果区间太长，这个假定可能受到破坏从而估计不精确。

19.2.2 生命表分析过程

1. 生命表分析基本过程

(1) 按 Analyze→Survival→Life Tables 顺序逐一单击鼠标，最后展开 Life Tables 生命表对话框，见图 19-1。



图 19-1 生命表主对话框

(2) Time 框：从左边的变量框中选择生存时间变量进入该框。生存时间可以是任何时间单位，如果在生存变量中有负数，在分析过程中会将其剔除。

(3) Display Time Intervals 栏：在该栏中确定时间的区间。默认单个值 0 作为第一个时间区间的开始点，读者在该选项的左侧框中输入所需要的最后一个时间区间的开始点，在右侧框中输入确定区间跨度的数值。例如，左侧框中输入“200”，在右侧框中输入“20”，这就表明最后一个区间的开始点为

200 个时间单位，从 0 至 200 每 20 个时间单位为一个区间。

(4) Status 框：选择状态变量进入 Status 框中，该变量用来标定删失和非删失状态。单击 Define Event 按钮，打开 Life Tables: Define Event for Status Variable 定义状态变量所发生事件对话框，见图 19-2。有两个选项：

① Single value，默认单个变量值为 0。对状态变量选择一个值，则系统只对状态变量为该值观测的生存时间进行分析，其他未选变量值的生存时间按删失值处理。例如，在状态变量中有 0、1、2、3 四种变量值，如果在该框中输入“2”，只对状态变量值为 2 的

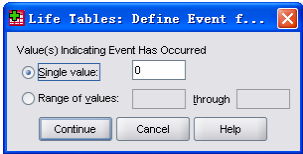


图 19-2 定义状态变量发生事件

观测进行生存时间分析，而忽略其他3种变量值观测的生存时间分析。

② Range of values through，指定状态变量值的范围。系统只对状态变量值在该范围内的观测的生存时间进行分析，其他值的生存时间按删失值处理。

(5) Factor 框：选择第一控制变量进入该框中。单击 Define Range 按钮，打开 Life Tables: Define Range for Factor 定义控制变量范围对话框，见图 19-3。不同处理方案导致不同的结果，选择第一控制变量进入 Factor 框中将不同的方案结果分别显示。

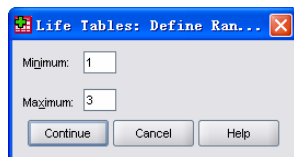


图 19-3 定义控制变量范围

Minimum 和 Maximum 框，输入最小和最大值，确定分析范围。不同的变量值代表不同的分层。其他未选变量值的生存时间按删失值处理，如果变量中有负值在分析过程中将被剔除。

(6) By Factor 框：选择第二控制变量进入本框中，并在 Life Tables: Define Event for Status Variable 对话框中确定分层的范围。第二分类变量中各分层将与第一分类变量中各分层相互结合，一一生成生命表细分组。

## 2. 生命表分析选择项

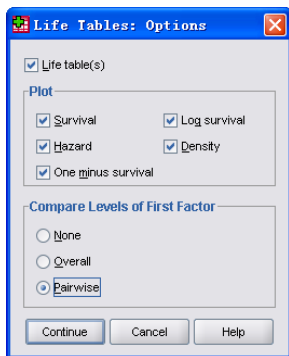


图 19-4 生命表选择项对话框

单击 Life Tables 对话框中的 Options 按钮，展开选择项对话框，见图 19-4。

(1) Life table(s) 不选择该项，将不生成生命表。

(2) 在 Plot 栏中选择生成的函数图形。

① Survival，以线性刻度生成累积生存函数图。

② Log survival，以对数刻度生成累积生存函数图。

③ Hazard，以线性刻度生成累积危险函数图。

④ Density，生成密度函数图。

⑤ One minus survival，生成 1 减累积生存函数图。

(3) 在 Compare Levels of First Factor 栏中，选择比较第一控制变量中各层间的显著性差异的方式，系统使用 Wilcoxon (Gehan) 检验。如果有第二控制变量，先以第二控制变量的各层进行分组，再对第二控制变量中各分组中的第一控制变量中各层之间进行比较。

① None，不进行各分层的比较。

② Overall，同时比较第一控制变量中各分层的差异。

③ Pairwise，配对比较第一控制变量中各层的差异。例如，在第一控制变量中有 3 个分层，将 1 对 2，2 对 3，1 对 3 进行比较，同时比较第一控制变量在各分层中的差异。

## 19.2.3 生命表分析实例

【例 1】有位科学工作者研究了饮食与肿瘤之间的关系，他将同种同龄的 90 只老鼠

分成3组，在环境相同的情况下，分别给予低脂饮食（Low fat）、饱和饮食（Saturated）和不饱和饮食（Unsaturated），并对每只老鼠的脚趾注射等量的肿瘤细胞，观测这些老鼠200天。在这段时间内，有些老鼠偶然死亡并且没有发现肿瘤，还有一些老鼠在观测结束时仍然没有肿瘤。以上资料源于《生存数据分析的统计方法》（Elisa T Lee 著，中国统计出版社）。

要求：做出不同喂养方式下的生存时间表，比较不同喂养方式下生存的时间是否有显著性差异，绘制各种函数图形。

1. 数据 data19-01 中的变量有：ID 实验鼠编号、FOOD 三种不同的喂养方式（编码与值标签是 1: low-fat, 2: saturated, 3: unsaturated）、STATUS 观测量状态（编码与值标签为 0: died 已死亡, 1: censored 删失数据）、TIME 生存的时间（天）。

## 2. 数据处理

(1) 按 Analyze→Survival→Life Tables 顺序逐一单击鼠标，最后展开 Life Tables 生命表对话框，如图 19-1 所示。

(2) 从左侧的变量列表框中选择 TIME 变量，送入右面的 Time 框中。

(3) 在 Display Time Intervals 栏中确定时间的区间。最后一个时间区间的开始点为 200，区间跨度为 20。

(4) 选择 STATUS 变量进入 Status 框中。单击 Define Event 按钮，打开 Life Tables: Define Event for Status Variable 对话框，见图 19-2，并在 Single value 框中输入 0。

(5) 选择 FOOD 变量进入 Factor 框，作为第一控制变量。单击 Define Range 按钮，打开 Life Tables: Define Range for Factor 对话框，见图 19-3，在 Minimum 和 Maximum 框中分别输入 1 和 3。

(6) 单击 Life Table 对话框中的 Options 按钮，展开 Life Tables: Options 对话框，见图 19-4。选中 Life table(s)复选项，选择 Plot 栏中所有函数图形选项，在 Compare Levels of First Factor 栏中选定 Pairwise 项。

(7) 单击 OK 按钮，提交计算。

## 3. 输出结果见表 19-1～表 19-5 和图 19-5。

表 19-1 为生命表，由于篇幅所限，只摘录低脂肪食物老鼠生命表，表中各项解释如下。(1)时间区间，(2)进入区间的例数，(3)活着退出的例数（删失例数），(4)暴露例数（危险人数），(5)死亡例数，(6)死亡率，(7)生存率，(8)累积生存率，(9)累积生存率标准误差，(10)死亡概率，(11)死亡概率标准误差，(12)危险率函数，(13)危险率标准误差。

表 19-2 为中位生存时间。其中低脂肪食物老鼠中位生存时间为 197.93（月）。用 Wilcoxon（Gehan）统计方法比较不同食物喂养老鼠导致癌症的生存时间。

表 19-3 为总体比较控制变量中不同水平的检验统计量。

表 19-4 为配对比较的检验统计量。

表 19-5 为平均得分，其中包括总例数、未删失例数、删失例数、删失百分比、平均

得分。

图 19-5(a)为生存函数对数图,图 19-5(b)为生存函数图,图 19-5(c)为 1-生存函数图,图 19-5(d)为密度函数图,图 19-5(e)为危险函数图。

表 19-1 低脂肪食物的老鼠生命表

Life Table													
First-order Controls	(1) Interval Start Time	(2) Number Entering Interval	(3) Number Withdrawing during Interval	(4) Number Exposed to Risk	(5) Number of Terminal Events	(6) Proportion Terminating	(7) Proportion Surviving	(8) Cumulative Proportion Surviving at End of Interval	(9) Std. Error of Cumulative Proportion Surviving at End of Interval	(10) Probability Density	(11) Std. Error of Probability Density	(12) Hazard Rate	(13) Std. Error of Hazard Rate
low-fat	0	30	0	30	0	.00	1.00	1.00	1.41	.000	.000	.00	.00
	20	30	0	30.000	0	.00	1.00	1.00	2.00	.000	.000	.00	.00
	40	30	0	30.000	2	.07	.93	.93	1.97	.003	.007	.00	.00
	60	28	0	28.000	4	.14	.86	.80	1.73	.007	.014	.01	.00
	80	24	0	24.000	3	.12	.88	.70	1.56	.005	.011	.01	.00
	100	21	0	21.000	1	.05	.95	.67	1.62	.002	.004	.00	.00
	120	20	0	20.000	0	.00	1.00	.67	1.75	.000	.000	.00	.00
	140	20	1	19.500	2	.10	.90	.60	1.62	.003	.009	.01	.00
	160	17	0	17.000	1	.06	.94	.56	1.62	.002	.005	.00	.00
	180	16	0	16.000	2	.12	.88	.49	1.45	.004	.010	.01	.00

表 19-2 中位生存时间

Median Survival Time	
First-order Controls	Med Time
low-fat	197.93
saturated	110.00
unsaturated	92.00

表 19-3 总体检验统计量

Overall Comparisons <sup>a</sup>		
Wilcoxon (Gehan) Statistic	df	Sig.
12.058	2	.002

a. Comparisons are exact.

表 19-4 配对比较检验统计量

Pairwise Comparisons <sup>a</sup>				
(I) food	(J) food	Wilcoxon (Gehan) Statistic	df	Sig.
1	2	3.676	1	.055
	3	11.913	1	.001
2	1	3.676	1	.055
	3	2.532	1	.112
3	1	11.913	1	.001
	2	2.532	1	.112

a. Comparisons are exact.

表 19-5 平均得分

Mean Scores					
Comparison Group	Total N	Uncensored	Censored	Percent Censored	Mean Score
1 vs.2 1	30	15	15	50.0%	8.400
2	30	23	7	23.3%	-8.400
1 vs.3 1	30	15	15	50.0%	15.400
3	30	30	0	.0%	-15.400
2 vs.3 2	30	23	7	23.3%	7.167
3	30	30	0	.0%	-7.167
1	30	15	15	50.0%	15.400
2	30	23	7	23.3%	7.167
3	30	30	0	.0%	-7.167
Overall Comparison					

由于选择的观测量较少,所以不太适合用生命表分析方法,但从生命表中可以看出 60 至 100 天内喂养低脂肪食物患肿瘤死亡率较高。3 种不同食物喂养,患癌症后所生存的时间,经过 Wilcoxon (Gehan)统计分析,它们之间存在显著性差异 ( $p<0.05$ )。在进行组间比较时,得到低脂 (low-fat) 食物和饱和 (saturated) 食物之间的生存时间有显著性差异,检验统计量为 11.913,自由度为 1,概率为 0.0006 ( $p<0.05$ )。通过观察统计图可

以直观地看出喂养不同食物的小鼠的生存时间有所不同，读者可以在 Chart Editor 窗口对密度函数和危险函数做出拟合线。

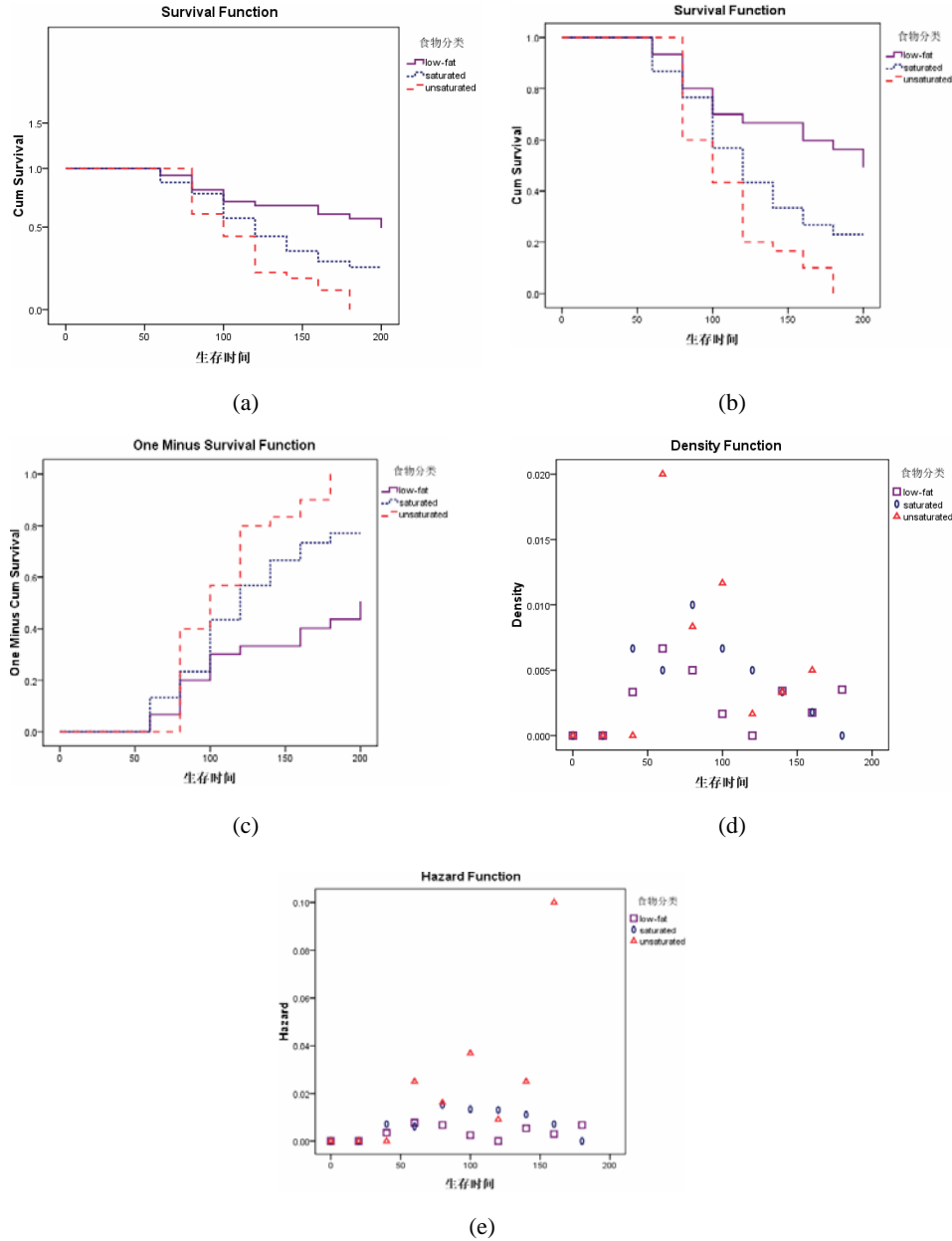


图 19-5 生存图形

## 19.3 Kaplan-Meier分析

### 19.3.1 Kaplan-Meier分析概述

对于 Kaplan 和 Meier (1958 年) 所提出的估计生存函数的乘积限 (Product-Limit, PL) 方法, 很多作者也把它称为生命表估计, 二者的差别是: PL 估计是基于一个个的数据, 而生命表估计是基于按区间分组数据。PL 估计可看成是生命表估计的特殊情形。

时间变量应是数值型。状态变量可以是二变量或分类变量, 发生的事件可以用一个正数值表示或用某个范围的连续数值表示。

生命表假设, 事件发生的概率仅依赖于时间。

### 19.3.2 Kaplan-Meier分析过程

#### 1. Kaplan-Meier 分析基本过程

(1) 按 Analyze→Survival→Kaplan-Meier 顺序逐一单击鼠标左键, 最后展开主对话框 Kaplan-Meier, 见图 19-6。

(2) Time 框: 从左侧的变量框中选择生存时间变量进入该框中, 生存的时间可以是任何时间单位, 如果在生存变量中有负数, 系统在分析过程中将其剔除。

(3) Status 框: 选择标定删失和非删失的状态变量进入 Status 框中。单击 Define Event 按钮, 打开 Kaplan-Meier: Define Event for Status Variable 对话框, 见图 19-7。在该对话框中选择要分析的状态, 系统只分析选定的状态下的生存时间数据, 其余按删失值处理。

① Single value 单个状态变量值。例如, 在状态变量中有 0、1、2、3 四种变量值, 如果在该框中输入“2”, 只对状态变量值为 2 的观测进行生存时间进行分析。

② Range of values through 指定状态变量值范围。例如, 在状态变量中有 0、1、2、3 四种值, 输入值为 1 和 3, 只对状态值为 1、2、3 的生存时间进行分析。

③ List of values, 变量值列表。例如, 在状态变量中有 0、1、2、3 四种变量值, 如果在该框中输入 1 和 3, 只分析状态值为 1 和 3 的生存时间。

(4) Factor 框: 选择控制变量进入该框。用短字符型或数值型的变量值代表不同水平。

(5) Strata 框: 选择分层变量进入本框, 即在控制变量中不同的处理方案内进行分层。该变量的值代表不同的分层。变量可以是短字符型也可以是数值型。

(6) Label Cases by 框: 选择标识观测量的变量进入本框, SPSS 将以列表方式用该变量值标出所有的观测量, 该变量可以是字符型, 其值可以是小于等于 20 个字母的字符串。

#### 2. 选择比较控制因素的统计方法

选择了控制变量后, 可以比较各个不同水平是否具有显著性差异。单击 Kaplan-Meier 对话框中的 Compare Factor 按钮, 展开选择比较控制因素统计方法对话框, 见图 19-8。

(1) 选择统计方法

① Log rank，即 Mantel-Haenszel 检验，又称时序检验，对所有的死亡时间赋予相等的权重，比较生存分布是否相同，它对于后期差别较为敏感。

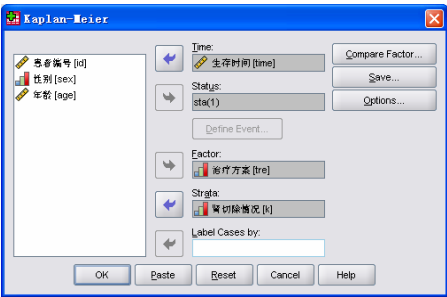


图 19-6 Kaplan-Meier 主对话框

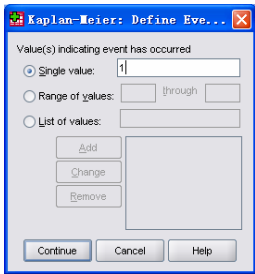


图 19-7 K-M 定义  
状态变量发生事件

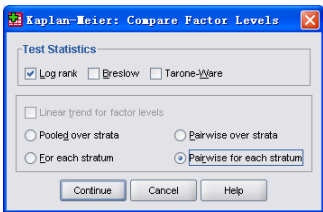


图 19-8 Kaplan-Meier 选择比  
较控制因素统计方法

② Breslow，对较早死亡时间赋予较大的权重，所以对于早期差别较为敏感。

③ Tarone-Ware，比较生存分布是否相同，当两个危险率函数曲线或生存曲线有交叉时，可以考虑使用 Tarone-Ware 检验。

(2) 选择比较的方式

① Linear trend for factor levels，因子水平的线性趋势，如果因子水平有自然顺序（如病情的早期、中期、晚期）时，选中该复选项，做趋势检验。

② Pooled over strata，合并比较所有因子水平下的生存时间，不进行配对比较。

③ For each stratum，如果选择了分层变量，在每层比较不同因子水平下的生存时间。

④ Pairwise over strata，以不同的配对方式比较每一对因子水平下的生存时间，如果选择了趋势检验，这种方法不能使用。

⑤ Pairwise for each stratum，如果选择了分层变量，在每层以不同的配对方式比较每一对因子水平下的生存时间。但选择了趋势检验，这种方法也不能使用。

3. 保存新的统计量

将运算中新的统计量保存到数据窗中，单击 Kaplan-Meier 对话框中的 Save 按钮，展开 Kaplan-Meier: Save New Variables 保存新变量对话框，见图 19-9。

① Survival，保存累积生存概率估测值，如果没有指定变量名，自动生成前缀带有“sur”的变量名，如“sur\_1”，“sur\_2”等。

② Standard error of survival 复选项，保存累积生存概率的标准误，如果没有指定变量名，自动生成前缀带有“se”的变量名，如“se\_1”，“se\_2”等。

③ Hazard 复选项，保存累积危险函数估测值，如果没有指定变量名，自动生成前缀带有“haz”的变量名，如“haz\_1”，“haz\_2”等。

④ Cumulative events 复选项，保存发生事件的累积频率，如果没有指定变量名，自

动生成前缀带有“cum”的变量名，如“cum\_1”，“cum\_2”等。

#### 4. Kaplan-Meier 分析选择项

用户根据需要选择一些统计量和图形。单击 Kaplan-Meier 对话框中的 Options 按钮，展开 Kaplan-Meier:Options 选择项对话框，见图 19-10。

##### (1) Statistics 统计量栏

① Survival table(s)，生成一个简化的生命表，它只包括乘积限生命表、标准误、累积频数、风险例数。如果清除该复选项，将不生成生命表，这样可以压缩输出的篇幅。

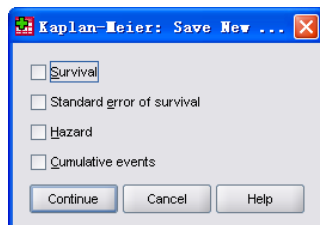


图 19-9 保存新变量对话框

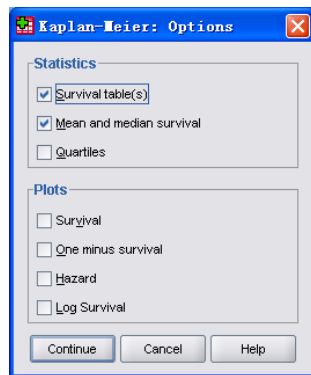


图 19-10 选择项对话框

##### ② Mean and median

survival，计算生存时间的均数、中位数及其标准误和置信区间。

③ Quartiles，输出结果显示生存时间的 25、50 和 75 分位数，以及它们的标准误。

##### (2) Plots 图形栏中选择生成的函数图形

① Survival，将生成线性刻度的累积生存函数图。

② One-minus survival，生成 1 减累积生存函数图。

③ Hazard，生成线性刻度的累积危险函数图。

④ Log Survival，生成对数刻度的累积生存函数图。

### 19.3.3 Kaplan-Meier 分析实例

【例 2】某医院对 58 例肾上腺样瘤患者在不同治疗研究中得到的数据，资料源于《生存数据分析的统计方法》(ELISA T. LEE 著，中国统计出版社)，数据编号 data19-02。

要求显示生存时间的均数和中位数，以及 25 分位数、50 分位数和 75 分位数；要求检验在切除肾脏条件下两种治疗方案的结果是否具有显著性差异。

#### 1. 数据

数据文件 data19-02 中的变量、变量标签、值、值标签为：id（患者编号）、sex（性别：1，男；2，女）；k（肾切除情况：0，未切；1，切除）、tre（治疗方案：1，化学与免疫疗法结合；2，其他方法）、time（生存时间：-99，未知）、sta（观测量的状态：0，删失数据；1，已死亡；9，未知）。

#### 2. 操作步骤

(1) 按 Analyze→Survival→Kaplan-Meier 顺序单击鼠标，展开如图 19-6 所示的对话框。

(2) 从左侧的变量框中选择 time 变量，送入 Time 框中。



- (3) 选择 sta 变量进入 Status 框中。单击 Define Event 按钮，打开 Kaplan-Meier: Define Event for Status Variable 对话框，见图 19-7，并在该框中 Single value 框中输入 1。
  - (4) 选择 tre 变量进入 Factor 框，作为控制变量，见图 19-6。
  - (5) 选择 k 变量进入 Strata 框，作为分层变量，见图 19-6。
  - (6) 单击 Compare Factor 按钮，展开 Kaplan-Meier: Compare Factor Levels 对话框，见图 19-8。选中 Log rank 项，同时选定 Pairwise for each stratum 项。
  - (7) 单击 Options 按钮，展开 Kaplan-Meier: Options 对话框，见图 19-10。选中 Mean and median survival 和 Quartiles 复选项。
  - (8) 单击 OK 按钮，提交计算。
3. 结果输出 见表 19-6～表 19-9

表 19-6 观测量删失情况

Case Processing Summary					
肾切除情况	治疗方案	Total N	N of Events	Censored	
				N	Percent
未切	化学与免疫法结合	7	7	0	.0%
	其他方法	3	3	0	.0%
	Overall	10	10	0	.0%
切除	化学与免疫法结合	29	25	4	13.8%
	其他方法	17	12	5	29.4%
	Overall	46	37	9	19.6%
Overall	Overall	56	47	9	16.1%

表 19-7 生存时间的平均值和中位数

Means and Medians for Survival Time									
肾切除情况  治疗方案		Mean <sup>a</sup>				Median			
		Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound			Lower Bound	Upper Bound
未切	化学与免疫法结合	12.571	2.034	8.585	16.558	12.000	3.928	4.301	19.699
	其他方法	8.000	.000	8.000	8.000	8.000			
	Overall	11.200	1.555	8.152	14.248	8.000	.949	6.141	9.859
切除	化学与免疫法结合	46.217	7.154	32.194	60.240	36.000	7.908	20.500	51.500
	其他方法	52.392	18.232	16.657	88.128	20.000	4.749	10.692	29.308
	Overall	47.414	7.698	32.326	62.503	30.000	6.982	16.316	43.684
	Overall	Overall	40.825	6.579	27.929	53.720	20.000	3.606	12.932

a. Estimation is limited to the largest survival time if it is censored.

表 19-8 生存时间的四分位数

Percentiles							
肾切除情况	治疗方案	25		50		75	
		Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
未切	化学与免疫法结合	17.000	2.315	12.000	3.928	8.000	1.793
	其他方法	8.000		8.000		8.000	
	Overall	15.000	3.795	8.000	.949	8.000	.791
切除	化学与免疫法结合	72.000	16.537	36.000	7.908	14.000	3.404
	其他方法	40.000	8.277	20.000	4.749	16.000	2.627
	Overall	68.000	12.163	30.000	6.982	14.000	2.962
Overall	Overall	52.000	14.250	20.000	3.606	10.000	1.614

表 19-6 为 KM 分析过程中观测量删失情况，KM 分析过程中将变量中的负数或缺失值剔除。Total N 总数、N of Events 未删失的例数、Censored N 删失数、Censored Percent 删失的百分比。

表 19-7 和表 19-8 为不同分层及不同处理情况生存描述性统计量，表 19-7 为生存时间的平均值和中位数以及它们 95% 的置信区间。表 19-8 为生存时间的四分位数，即 25%、

50%、75%的数值。

表 19-9 为 Log Rank 检验统计量，在分层变量为 0 值时，对控制变量不同的水平做时序检验（Log Rank）。

表 19-9 Log Rank 检验统计量

Pairwise Comparisons						
Log Rank (Mantel-Cox)	肾切除情况	治疗方案	化学与免疫法结合		其他方法	
			Chi-Square	Sig.	Chi-Square	Sig.
	未切	化学与免疫法结合	2.440	.118	2.440	.118
		其他方法				
切除	化学与免疫法结合	.110	.741	.110	.741	
	其他方法					

统计结果表明，对 58 名肾上腺样瘤的治疗中，患者的肾脏切除或不切除，化学与免疫结合的疗法同其他疗法，在延长患者生存时间上没有显著性差别。在肾脏切除的情况下，Log Rank 检验统计量为 2.44 ( $p>0.05$ )；在肾脏未切除的情况下，Log Rank 检验统计量为 0.11 ( $p>0.05$ )。

## 19.4 Cox Regression 风险比例模型分析

### 19.4.1 Cox Regression 分析概述

在 Cox 模型中，生存时间或恢复时间常作因变量，而与生存时间有关的一组变量作为自变量，即预后变量或协变量。

时间变量应是数值型。状态变量可以是分类或连续型变量。如果是分类变量，应是哑变量或用指示编码。分层变量为分类变量，可用整数或短字符串编码。自变量（协变量）可以是分类型或连续型变量。预后变量可是连续变量或离散变量。连续型自变量可以直接用在方程里，若是离散型变量，必须编码成指示变量才能参与分析。指示变量可以在 Categorical Covariates 对话框重新编码。关于指示变量的编码方式见第 11 章。

在拟合 Cox 模型之前，可以通过计算变量之间的相关系数来查明与因变量显著相关的变量，对数据的质量进行检查，然后结合专业知识拟合模型。应注意没有进入模型中的因素不一定是无关的因子，进入模型中的因子也不一定就是相关因子。

比例风险假设为，从一个事件到另一个事件的风险比例不随时间而变化。

### 19.4.2 Cox Regression 分析过程

#### 1. Cox Regression 分析基本过程

(1) 按 Analyze→Survival→Cox Regression 顺序逐一单击鼠标左键，最后展开 Cox 回归风险比例模型主对话框，见图 19-11。

(2) Time 框：从左侧的变量列表中选择生存时间变量进入该框中，生存的时间可以

是任何时间单位的连续型变量。在分析中自动剔除生存变量中的负数。

(3) **Status 框**: 选择标定删失和非删失状态的状态变量进入 **Status** 框中。单击 **Define Event** 按钮, 打开 **Define Event for Status Variable** 对话框, 见图 19-7。选择要分析的状态。具体方法见 19.3.2 节。

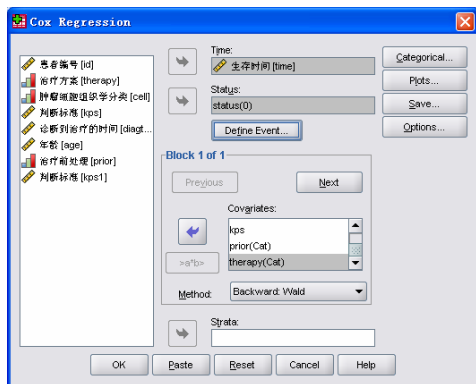


图 19-11 Cox 回归模型主对话框

(4) **Covariates 框**: 从变量列表框中选定一个或多个协变量进入本框, 协变量可以是连续型或分类型变量。

通过使用 **Previous** 与 **Next** 按钮, 指定不同的协变量组, 单击 **Next** 按钮进入下一个协变量组, 单击 **Previous** 按钮退回到上一个协变量组。

如果考虑协变量间的交互作用, 在变量表中选择有交互作用的变量, 单击 **a\*b** 按钮, 形成交互作用项进入 **Covariates** 框。

(5) **Method 框**: 在该框的列表中选择协变量进入回归模型的方式, 共 7 种。

① **Enter**, 强行进入法, 同一组中的协变量,

一次性地全部进入回归方程。

② **Forward Condition**, 变量经过条件似然检验确定是否进入回归方程的向前选择法。

③ **Forward LR**, 变量经过似然率检验确定是否进入模型的向前选择法。

④ **Forward Wald**, 变量经过沃德检验确定是否进入模型的向前选择法。

⑤ **Backward Condition**, 变量经过条件似然检验确定是否从模型剔除的向后选择法。

⑥ **Backward LR**, 变量经过似然比检验确定是否从模型中剔除的向后消去法。

⑦ **Backward Wald**, 变量经过沃德检验确定是否从模型中剔除的向后消去法。

一般来说, 使用向后消去法可以减少漏掉潜在的有价值的预测因子。如果至少有一个协变量进入模型, 可以使用向前选择法。

(6) **Strata 框**: 选定分层变量进入本框, SPSS 根据分层变量将数据细分组, 然后在每个分组数据的基础上生成各自的风险函数。分层变量应是分类变量。

## 2. 分类变量的编码

在主对话框中单击 **Categorical** 按钮, 展开 **Cox Regression: Define Categorical Covariates** 定义分类协变量对话框, 如图 19-12 所示。对于数值型的分类变量需要在本对话框中重新编码, 新的编码变量名后就标注 “Cat”。

(1) **Covariates 框**中为所有在主对话框中选定的协变量。从中选择要编码的数值型分类自变量送入 **Categorical Covariate** 框中。

(2) 在分类协变量框中选择一个变量, 在 **Change Contrast** 栏中选择一个对比类型和

对比类。可以选择下列对比类型：

① **Deviation** 离差对比。预测变量中每个分类效应与总效应比较。

② **Simple** 简单对比。预测变量的每类与参照类比较。可选择 **First** 第一类或 **Last** 最后一类作为参考类。

③ **Difference** 差别对比，除第一类外，预测变量的每一类都与该类前面的各类的平均效应相比较。又称为反赫尔默特对比。

④ **Helmert** 赫尔默特对比，除最后一类外，预测变量的每类与后面各类的平均效应相比较。

⑤ **Repeated** 重复对比，除第一类外，预测变量的每个分类都与它前面的分类比较。

⑥ **Polynomial** 正交多项式对比，只能用于数值型分类变量。且假定各类间有相等的空间。

⑦ **Indicator** 指示对比，指明类代表信息的有无，可选 **First** 或 **Last** 作为参考类。

⑧ **Reference** 离差对比、简单对比和指示对比中，读者可以去除默认的参照分类，可以选择第一或最后一个分类作为默认分类。

完成选择后，单击 **Change** 按钮，确定这些设置。

### 3. 生成图形

单击 **Cox Regression** 对话框中的 **Plots** 按钮，展开 **Cox Regression Plots** 图形对话框，见图 19-13。在图形对话框中，读者可以获得以下图形。如果有时间相依性协变量，将不能生成图形。

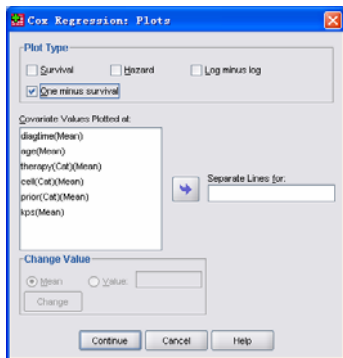


图 19-13 Cox 模型图形对话框

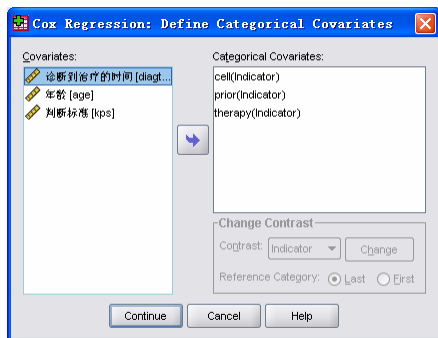


图 19-12 定义分类协变量对话框

#### (1) Plot Type 栏：

① **Survival** 生成线性刻度的累积生存函数图形。

② **Hazard** 生成线性刻度的累积危险函数图形。

③ **Log minus log** 生成经过  $\ln(-\ln)$  转换之后的累积生存估计值的图形。

④ **One-minus survival** 生成 1 减累积生存函数图。

(2) **Covariate Value Plotted at** 框：在默认状态下，以模型中对比变量和协变量的均值绘制函数图形，也就是在本框中单击 **Mean** 选项，再单击 **Change** 按钮。如果以对比变量和协变量其他数值绘制函数图形，选中本框中的一个或多个协变量，然后在 **Change Value**

框中单击 **Value** 选项，并在其参数框中输入数值，最后单击 **Change** 按钮，SPSS 根据读者指定的协变量值，绘制其危险函数和生存函数。

(3) **Separate Lines for** 框: 选择一个分类协变量进入本框, 系统按变量值将数据分成两个或多个分组, 对各分组分别绘制函数图。如果指定了层变量, 则每层绘制一个图。

#### 4. 保存新的统计量

主对话框中单击 **Save** 按钮, 展开如图 19-14 的保存新变量对话框。选择要保存在数据窗中的分析结果, 作为新变量有生存变量和诊断变量。

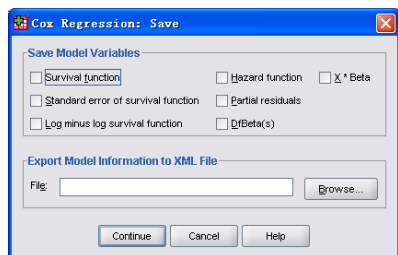


图 19-14 保存 Cox 模型新变量对话框

(1) **Survival** 栏指定生成的生存变量:

① **Function** 保存生存函数估测值, 自动生成的变量名前缀为“sur”, 如 sur\_1、sur\_2 等。

② **Standard error** 生存函数估测值的标准误。自动生成的变量名前缀为“se”。

③ **Log minus log** 经对数—对数转换的生存函数估测值。新变量名前缀为“lml”。

(2) **Diagnostics** 栏指定生成的诊断变量:

① **Hazard function** 累积危险函数估测值。自动生成的变量名前缀为“haz”。

② **Partial residuals** 生成对生存时间的偏残差, 用以检验比例危险的假设, SPSS 为最终模型中每个协变量保存一个偏残差变量。在模型中至少含有一个协变量才能生成偏残差。自动生成的变量名前缀为“pr”, 如 pr1\_1、pr1\_2、pr2\_1、pr2\_2 等。

③ **DfBeta(s)** 每个观测量从模型拟合中被剔除时, 标准化回归系数的变化量。模型中至少含有一个协变量才能生成标准化回归系数变化量变量。新变量名前缀为“dfb”。

(3) **X\*Beta** 线性预测因素得分。它是平均中心协变量值与其相对应的每个观测量参数估计值的乘积和。新变量名前缀为“xbe”。

#### 5. Cox Regression 分析选项

在主对话框中单击 **Options** 按钮, 展开选项对话框, 选择统计和输出方式, 见图 19-15。

(1) **Model Statistics** 统计量栏

① **CI for exp(B)** 设置相对危险估计值的置信区间, 常用的有 90%、95% 和 99%。

② **Correlation of estimate** 显示回归系数估计值的相关系数矩阵。

③ **Display model information** 栏: 对当前模型显示对数似然统计量、似然比统计量和总体卡方值。对模型中的变量, 显示参数估计值及其标准误, Wald 统计量。对已剔除出模型的变量, 显示记分检验统计量和残差卡方值。

- **At each step** 在逐步回归的每一步显示上述全部统计量。

- **At last step** 显示逐步回归最后一步进入模型的协变量和最后模型的统计量。

(2) **Probability for Stepwise** 栏: 如果选择了逐步回归法, 还应该在 **Entry\_Removal\_**

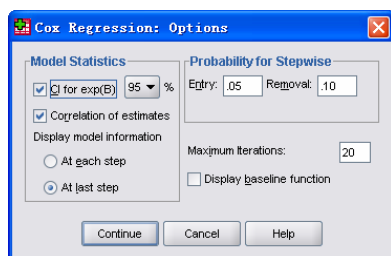


图 19-15 Cox 模型选项对话框

框中指定协变量进入或剔除出模型的概率。进入、默认的概率分别为 0.05 和 0.10。注意，进入概率值应该小于剔除概率值，否则模型中将没有变量。

(3) Maximum Iterations 框：为模型指定最大迭代数。用 Newton-Raphson 方法计算参数估计值时，如果达到最大迭代数，迭代过程将停止。

(4) Display baseline function，生成基准危险函数、协变量均值生存和危险函数表。若有分层变量，则每层生成独立表格。若指定了时间相依性协变量，不能激活该选项。

### 19.4.3 Cox Regression分析实例

【例 3】下面是一组 137 位肺癌患者生存时间的数据。该数据来自《SAS/STAT guide for personal computers》，用 Cox 回归模型辨认预测因素。

1. 数据在 data19-03 文件中。需要说明的变量：diagtime 诊断到治疗的时间、time 生存时间、prior 治疗前处理（0：经过处理；1：未经处理）、therapy 治疗方案（1：标准方法；2：实验方法）、status 病人状态（0：死亡；1 删失数据）cell 肺癌细胞组织学分类（1：鳞癌；2：小细胞肺癌；3：腺癌；4：大细胞肺癌）、kps 判断标准（≤30 住院治疗；30~60 住院和家庭治疗；>60 家庭治疗）。

#### 2. 操作步骤

(1) 按 Analyze→Survival→Cox Regression 顺序单击鼠标，展开如图 19-11 所示的对话框。

(2) 从左面的变量表中选择 time 变量，送入右面的 Time 框中。

(3) 选择 status 变量送入 Status 框中。单击 Define Event 按钮，在打开的对话框 Single value 编辑区中输入 0。

(4) 选择 age、cell、diagtime、kps、prior、therapy 作为协变量送入 Covariates 框。

(5) 在 Method 框中选择 Backward: Wald 项。

(6) 单击 Categorical 按钮，展开相应的对话框。选择 cell、prior、therapy 变量进入 Categorical 框中。选中这三个变量，使它们的对比方式均为 Indicator，其中 cell 变量参考类为 first，其他两个分类变量的参考类为 Last。

(7) 单击 Options 按钮，展开相应的对话框，见图 19-15。选 Correlation of estimate，在 Display model 栏内选择 At last step 项，在 Entry\_Removal\_框内分别输入 0.05 和 0.10，在 Maximum Iterations 框输入 20。

(8) 单击 OK 按钮，提交计算。

#### 3. 输出结果

表 19-10 是对数据处理说明。即观测量总数 128、有缺失值的观测量数 9、带有负生存时间的观测量数 0、在分层中删失观测量数 0、去除的观测量总数 0、用于统计分析的观测量数 137。以及它们占总观测量的百分比。

表 19-11 是各变量值编码。cell 分类变量，以该变量中的第一分类（即 squamous）

作为参照分类（编码 0、0、0）。(1)代表 small 类，(2)代表 adeno 类，(3)代表 large 类。

表 19-12 第一步全模型与最后一步模型对系数检验的对数似然比值、总体分数的卡方检验、从前一步到本步变化量的卡方检验等。

表 19-13 中使用向后剔除拟合的第一步和最后一步的统计量和沃德检验，Step1 第一步全部指定的协变量进入模型，但 Wald 检验说明只有 cell、kps 两变量对模型贡献显著 Step5 第 5 步说明经过一步步剔除对模型没有统计意义的协变量，最后剩下 cell、kps。

表 19-10 观测量处理表

Case Processing Summary		
	N	Percent
Cases available in analysis	Event <sup>a</sup>	128 93.4%
	Censored	9 6.6%
	Total	137 100.0%
Cases dropped	Cases with missing values	0 .0%
	Cases with negative time	0 .0%
	Censored cases before the earliest event in a stratum	0 .0%
	Total	0 .0%
	Total	137 100.0%

a. Dependent Variable: 生存时间

表 19-11 各变量值编码

Categorical Variable Codings <sup>a,d,e</sup>				
	Frequency	(1) <sup>a</sup>	(2)	(3)
therapy <sup>a</sup>	1=标准方法	69	1	
	2=实验方法	68	0	
cell <sup>a</sup>	1=鳞癌	35	0	0
	2=小细胞肺癌	48	1	0
	3=腺癌	27	0	1
	4=大细胞肺癌	27	0	0
prior <sup>a</sup>	0=经过处理	97	1	
	1=未经处理	40	0	

- a. Indicator Parameter Coding  
b. The (0,1) variable has been recoded, so its coefficients will not be the same as for indicator (0,1) coding.  
c. Category variable: therapy (治疗方案)  
d. Category variable: cell (肺癌细胞组织学分类)  
e. Category variable: prior (治疗前处理)

表 19-12 模型系数综合检验

Omnibus Tests of Model Coefficients <sup>a,c</sup>									
Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block	
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df
1 <sup>a</sup>	950.359	65.917	8	.000	61.409	8	.000	61.409	8
5	952.997	63.219	4	.000				58.771	4

- a. Variable(s) Entered at Step Number 1: age cell diagtime kps prior therapy  
b. Beginning Block Number 0, Initial Log Likelihood function: -2 Log likelihood: 1011.768  
c. Beginning Block Number 1. Method = Backward Stepwise (Valid)

表 19-13 进入方程变量的统计量

Variables in the Equation						
	B	SE	Wald	df	Sig.	Exp (B)
Step 1						
age	-.009	.009	.844	1	.358	.991
cell			17.916	3	.000	
cell(1)	.856	.275	9.687	1	.002	2.355
cell(2)	1.188	.301	15.610	1	.000	3.281
cell(3)	.400	.283	1.999	1	.157	1.491
diagtime	.000	.009	.000	1	.992	1.000
kps	-.033	.006	35.112	1	.000	.968
prior	-.072	.232	.097	1	.755	.930
therapy	-.290	.207	1.958	1	.162	.748
Step 5						
cell			17.080	3	.001	
cell(1)	.712	.253	7.939	1	.005	2.038
cell(2)	1.151	.293	15.441	1	.000	3.161
cell(3)	.325	.277	1.381	1	.240	1.384
kps	-.031	.005	35.612	1	.000	.970

表 19-14 模型外变量的统计量

Variables not in the Equation <sup>a</sup>			
Step 5	Score	df	Sig.
age	.424	1	.515
diagtime	.165	1	.684
prior	.248	1	.618
therapy	1.650	1	.199

a. Residual Chi Square = 2.675 with 4 df Sig. = .614

表 19-15 回归系数相关矩阵

Correlation Matrix of Regression Coefficients			
	cell(1)	cell(2)	cell(3)
cell(2)	.559		
cell(3)	.473	.581	
kps	.097	.257	.105

表 19-16 协变量均值

Covariate Means	
	Mean
therapy	.504
cell(1)	.255
cell(2)	.350
cell(3)	.197
kps	58.569
diagtime	8.774
age	58.307
prior	.708

表 19-14 为拟合结束时，未进入模型变量的统计量。检验结果 Sig 都大于 0.05，表明对模型无统计意义的变量都没有进入模型。

表 19-15 为回归系数的相关矩阵。相关系数均不大,说明进入模型的变量之间相互基本是独立的,共线性问题不明显。

表 19-16 为协变量均值。

从以上统计结果表明, kps 和 cell 变量对模型有显著性意义。kps 变量相对危险度为 0.970, 回归系数为 -0.031, 说明 kps 变量取值越大, 生存时间越长。在 cell 变量中, adeno 和 small 分类与 squamous 分类相比具有显著性, 而 large 与 squamous 相比不具有显著性差异。adeno 的回归系数为 1.151, 相对危险度为 3.161; small 回归系数为 0.712, 相对危险度为 2.038; large 回归系数为 0.325, 相对危险度为 1.384; 所以鳞癌细胞肺癌患者生存时间最长, 其次大细胞肺癌患者, 再次小细胞肺癌患者, 腺癌细胞肺癌患者生存时间最短。

## 习 题 19

1. 什么是生命表和 Cox 模型?

2. data19-04 数据为 3 期和 4 期黑瘤患者的数据, 其中: id 变量为编号, age 变量为年龄, sex 变量为性别 (1, 男, 2, 女), survtime 变量为生存时间, survstatus 变量为生存状态 (0, 死亡; 1, 删失), stage 变量为肿瘤级别。计算时间间隔为 5 个月的不同肿瘤级别生命表。(本数据来源:《生存数据分析的统计方法》Elisa Lee 著, 陈家鼎等译, 中国统计出版社, 北京 1998 年 4 月第 1 版)

3. data19-05 数据收集 63 例患者的生存时间、结局及影响因素。各变量的含义见表 19-17。请用 Cox 模型进行预后分析。(本数据来源于《医学统计学》孙振球主编, 人民卫生出版社, 北京 2002 年 8 月第 1 版)

表 19-17 某恶性肿瘤的影响因素及量化值

变 量	意 义	值标签 (或单位)
X0	编号	
X1	年龄	岁
X2	性别	1, 男, 2, 女
X3	组织学类型	0, 低分化, 1, 高分化
X4	治疗方式	0, 新方法, 1, 传统方法
X5	淋巴结是否转移	0, 否, 1, 是
X6	肿瘤的浸润程度	0, 未突破浆膜, 1, 突破浆膜
t	生存时间	月
Y	患者结局	0, 死亡, 1, 截尾



# 第 20 章 生成统计图形

## 20.1 概 述

统计图是用点的位置、线段的升降、直条的长短或面积的大小等方法表达统计资料的一种形式，其特点是简明生动、形象具体和通俗易懂。

SPSS 制图功能很强，能绘制许多种统计图形，这些图形可以由各种统计分析过程产生，也可以直接从 Graphs 图形菜单中的一系列图形选项直接产生一部分。Graphs 菜单提供三类图形：Chart Builder、一般统计图形和交互式图形。Chart Builder 实际上是 SPSS 为读者提供一些预设的图形库，读者按 Graphs→Chart Builder 顺序单击鼠标打开相应的对话框，然后根据对话框中的提示通过拖曳变量，非常快捷地生成图形。本章主要介绍一般统计图形和交互式图形。

SPSS 系统直接从当前数据窗口中读取指定数据而生成图形，数据影响图形的生成，因而在生成图形之前需要完成以下几个步骤。

### 1. 建立数据文件

打开数据窗口，录入有关数据。现有我国 12 座大中城市 1985—1994 年每月平均气温的数据文件，数据文件结构是以各个城市的月平均气温、年代和月份作变量。本资料来源于 1986—1995 年《中国统计年鉴》（中国统计出版社），数据文件编号 data20-01。

### 2. 制定数据文件结构

数据文件结构往往决定着图形的类型，同样是来源于同一个资料，可以做成不同的数据文件结构。例如，data20-01 数据文件结构可以生成 1985—1994 年某个城市十二个月份平均气温图，但不能生成 1985—1994 年某个月份各城市平均气温图，必须改变 data20-01 数据文件结构，以每个月份的平均气温、年代和城市作变量重新制作数据文件结构。数据文件编号 data20-02。

### 3. 调整数据文件结构

为了制图需要对已有的数据文件结构进行一些调整，就 data20-02 文件而言，若生成 1985—1994 年上海 4 月、5 月和 6 月的平均气温图，就必须将上海的数据单独生成一些变量。将凡是上海的数据复制，然后在当前的数据文件中粘贴数据形成有关上海气温的变量，也可以在一个新数据文件中粘贴这些数据形成有关的变量，数据文件编号 data20-03。为了区别已有的变量，将有关上海变量的变量名加上“sh”。

## 20.2 条形图和 3-D 条形图

条形图 (Bar Charts) 是利用相同宽度条形的长短或高低表现统计数据大小或变动的统计图, 条形图还有其他别名, 条形图横排称为带形图, 纵排又称柱形图。平面条形图只能显示两个变量, 而 3-D 条形图可以同时显示三个变量。

### 20.2.1 选择图形类型

读者在条形图生成条形图时, 首先选择图形类型。

按 Graphs→Legacy Dialogs→Bar 顺序单击鼠标, 打开 Bar Charts 条形图主对话框, 见图 20-1。

按 Graphs→Legacy Dialogs→3-D Bar 顺序单击鼠标, 打开 3-D Bar Charts 条形图主对话框, 见图 20-2。

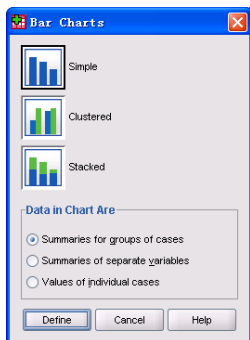


图 20-1 条形图主对话框

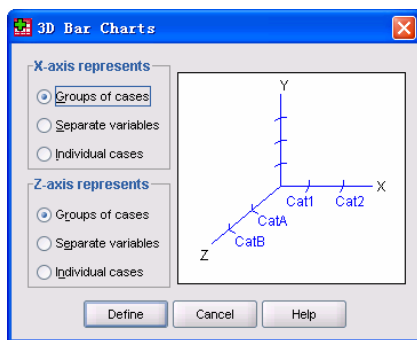


图 20-2 3-D 条形图主对话框

#### 1. 条形图式

- (1) Simple 简单条形图。以若干平行且等宽的矩形表现数量对比关系, 条间有间隙。
- (2) Clustered 分类条形图。由两条或两个以上条组成一组的条形图。
- (3) Stacked 分段条形图, 又称堆栈条形图。它是以条形的全长代表某个变量的整体, 条内的各分段长短代表各组成部分在整体中所占比例, 每一段用不同线条或颜色表示。

#### 2. 数据描述模式

- (1) Summaries for groups of cases 观测测量分组描述模式。每一组观测测量生成一个简单、分类、分段图形。
- (2) Summaries of separate variables 变量描述模式。每个变量生成一个图形, 这种模式至少要选择两个或两个以上、相同或不同的变量。
- (3) Values of individual cases 观测值模式, 每一观测值生成一个图形。

在 3D 数据描述对话框中, 数据模式的名称分别为 Groups of cases、Separate variables

和 Individual cases，其功能与上述三个选项相对应。

### 20.2.2 观测量分组描述简单条形图

读取 data20-04 数据文件。在条形图主对话框中选择 Simple，在 Data in Chart Are 栏内选择 Summaries for groups of cases，单击 Define 按钮，展开 Define Simple Bar: Summaries for Groups of Cases 观测量分类模式简单条形图对话框，见图 20-3。

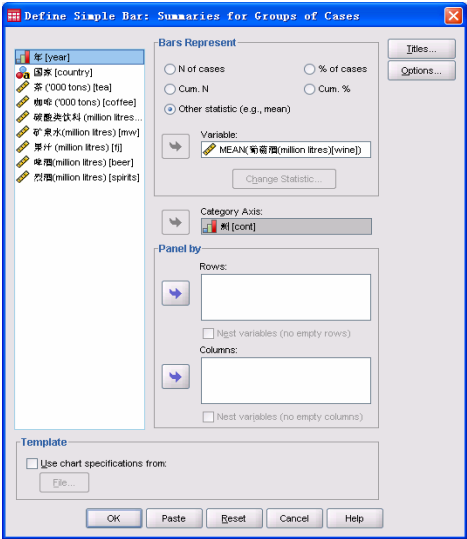


图 20-3 观测量分类模式简单条形图对话框

在本对话框中定义图形参数，

(1) Category Axis 框，设置分类轴变量。

在变量表中选择 cont 为分类轴变量，送入 Category Axis 框中。默认的分类轴是横轴。

分类轴上各变量值的排列位置，是由分类变量中变量值的大小和字母顺序所确定的，数值最小或字母顺序最靠前的变量值排在分类轴的最左端，相反则排在最右端。

在变量列表框中选择纵轴变量移入 Variable 框内，本例选择 wine。

(2) Bars Represent 栏，选择条图表达的统计量，共两大类：一类是对分类变量的描述；另一类是对其他变量的描述。

① 分类变量的计数函数，表达某一变量值的：

- N of cases 观测量计数。
- Cum.N 累积频数。
- % of cases 观测量数占总数的百分比。
- Cum % 累积百分数。

② Other statistic 其他综合统计函数

Variable 框中所显示的是统计函数表达式：Mean[wine]，Mean 为统计函数，wine 为统计函数的自变量，默认条长表示葡萄酒产量均值。倘若选择其他统计函数，单击 Change Summary 按钮，展开如图 20-4 所示的 Summary Function 综合函数选择对话框，在对话框内选择 Variable 框中表达式的统计函数部分。统计函数共 4 组 18 个选项。

第一组包括 10 个统计函数选项：

- Mean of values 算术平均数。
- Median of values 中位数。
- Mode of values 众数。

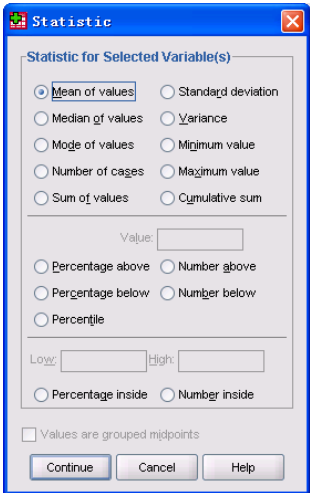


图 20-4 综合函数选择对话框

- Number of cases 不含缺失值的观测量数目。
- Sum of values 总和。
- Standard deviation 标准差。
- Variance 方差。
- Minimum value 最小值。
- Maximum value 最大值。
- Cumulative sum 累积总和。

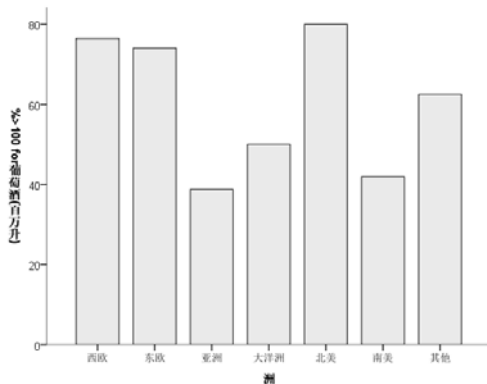
第二组包括 5 个统计函数项：

- Percentage above 大于指定参数的观测量数目占总数的百分比。
- Percentage below 小于指定参数的观测量数目占变量值总数的百分比。
- Number above 大于指定参数的观测量数。
- Number below 小于指定参数的观测量数。
- Percentile 百分位数。

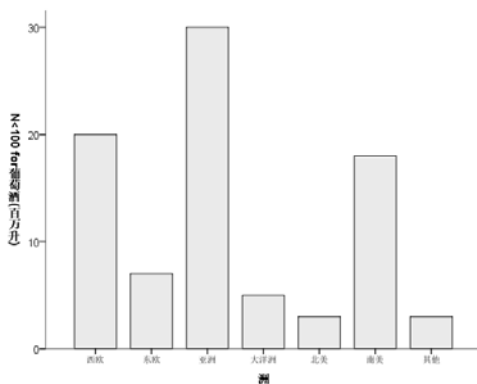
本组的统计函数项与 Value 框中的参数有关联，确定了统计函数选项后，在 Value 框内指定一个参数，Value 框只能录入 12 个字符。

图 20-5(a)以 cont 变量作为分类变量，在 Bar Represent 栏选择 Other summary function 项，将 wine 变量选入 Variable 框内，单击 Change Summary 按钮，在 Summary Function 框中选择 Percentage above 项，并在其参数框中输入 100。生成 1988—1992 年各洲葡萄酒产量大于 100 万升国家占该地区国家总数的百分比条形图。

图 20-5(b)与(a)不同的是，在 Summary Function 框中选择 Number below，并在参数框中输入 100。生成 1988—1992 年各洲葡萄酒产量小于 100 万升的国家数量对比条形图。



(a)



(b)

图 20-5 例图

第三组包括两个函数选项：

- Percentage inside 落在 Low 和 High 框参数范围内的观测量数占总数的百分比。

- Number inside 落在 Low 和 High 框参数范围内的观测量数目。

选择统计函数项后, 在 Low、High 框内指定下限、上限值, 可以是小于等于 7 个字符的值, 两个参数在自变量值的范围内, 且  $Low < High$ 。生成图形的分类轴包括 Low、High 两个点; 生成图形剔除自变量中的缺失值。

第四组仅一个选项, Values are grouped midpoints 变量值以中点分组, 选择了 Median of values 和 Percentile 项, 该选项有效。选中此项, 计算中位数和百分位数。



图 20-6 图题对话框

### (3) Title 图题和注释

单击 Titles 按钮, 出现 Titles 图形标题对话框, 如图 20-6 所示。Titles 对话框分为三个框: Title 图形标题框, Subtitle 图形子题框和 Footnote 注释框。具体操作参见图形编辑窗口的 Titles 命令。

(4) 单击 Options 按钮, 打开如图 20-7 所示 Options 对话框, 选择缺失值处理和误差条图的显示方式。

### ① Missing Values 栏, 选择缺失处理方式

- Exclude cases listwise, 在 Bars Represent 框所指定的各个变量中, 如果某个观测量在一变量中有缺失值,

那么剔除整个观测量。

- Exclude cases variable by variable 在 Bars Represent 框所指定的变量中存在缺失值, 仅剔除这个变量的缺失值。
- Display groups defined by missing values, 显示缺失值所定义的组。

② Display chart with case labels 在图形中显示观测量的标签值。

③ Display error bars 选择误差条图所表达的统计量。

- Confidence intervals 置信区间, 在 Level(%)框输入需要的水平值。
- Standard error 标准误, 在 Multiplier 框中根据输入标准误的倍数。
- Standard deviation 标准差, 在 Multiplier 框中根据输入标准差的倍数。

### (5) Template 图形模板格式

为简化操作, 可以套用已经做好的图形模板。选择 Template 框中的 Use chart specifications form 项后, 单击 File 按钮, 出现 Use template from file 应用模板格式文件对话框, 指定模板文件。新生成的图形其大小、比例、小数位数、字形、字体以及图题的位置等都自动转换成模板格式。

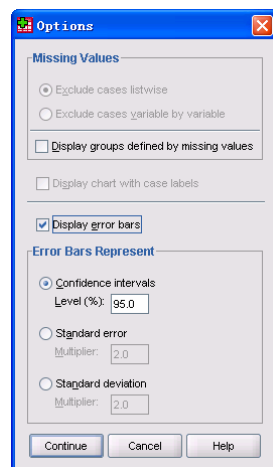


图 20-7 缺失值处理对话框

### 20.2.3 变量模式简单条形图

读取 data20-05 数据文件。在条形图主对话框中选择 Simple 和 Summaries of Separate Variables 项, 单击 Define 按钮, 展开 Define Simple Bar: Summaries of Separate Variables 变量模式简单条形图对话框, 见图 20-8。

(1) Bar Represent 框, 在条形图表达统计量框中至少要有两个或两个以上的变量。所选的变量可以是不同变量, 也可以是相同的变量。如果想要改变每个变量统计函数, 先将光标移至要改变的统计函数表达式, 单击 Change Summary 按钮, 出现 Summary Function 对话框, 见图 20-8。

变量在 Bar Represent 框中上下位置, 决定着这些被选变量在分类轴上从左向右排列的顺序, 在 Bar Represent 框中最上端的变量, 排在分类轴上的最左端。如果所选变量相同, 而统计函数不同, 分类轴标记为函数名; 如果所选变量不同, 而统计函数相同, 分类轴标记为变量名。

本小节选择收缩压(sp)和舒张压(dp)变量进入 Bar Represent 框。

(2) Panel by 栏, 生成群组图形。即由若干按照一定方式排列的小图形组成。

① Rows 框: 确定横向排列图形的变量, 将性别(sex)变量选入该框内。

② Columns 框: 确定纵向排列图形的变量, 将年龄(age)变量选入该框内。

单击 OK 按钮, 生成不同年龄不同性别血压变化图, 见图 20-9。

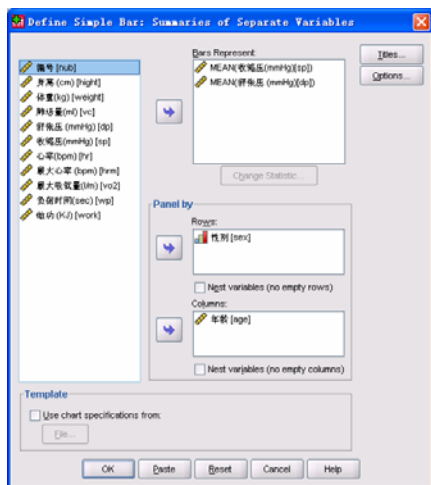


图 20-8 变量模式简单条形图对话框

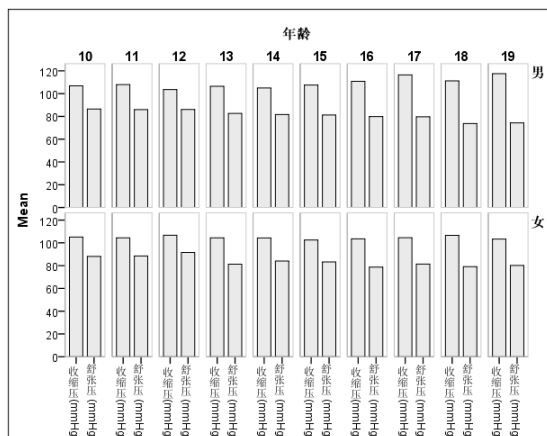


图 20-9 不同年龄不同性别血压变化图

### 20.2.4 观测量分组模式分段条形图

1. 读取 data20-04 数据文件。

2. 在条形图主对话框中选择 Stacked 和 Summaries for groups of cases 项, 单击 Define 按钮, 展开观测量分类模式分段条形图对话框, 见图 20-10。

- 3. Bars Represent 栏中选择 Other summary function 项，选 cc 变量送入 Variable 栏。
  - 4. Category Axis 选择 cont 变量作为分类轴变量送入 Category Axis 框中。
  - 5. Define Stacks by 在变量框中选择 year 作为分段变量，送入 Define Stacks by 框中。
- 分段是以分段变量中各变量值的数字或字母顺序排列，数值小或字母顺序靠前的变量值在条形图的下端，大的或靠后的在条形图的上端。

单击 OK 按钮，生成 1988—1992 年各洲每年碳酸盐和浓缩饮料平均产量，见图 20-11。

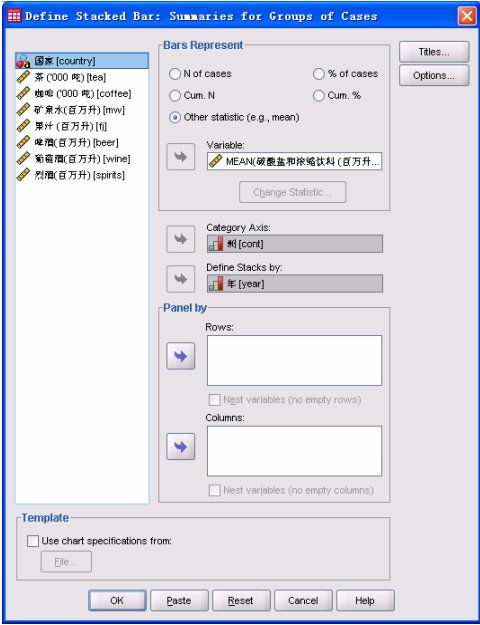


图 20-10 观测量分类模式分段条形图对话框

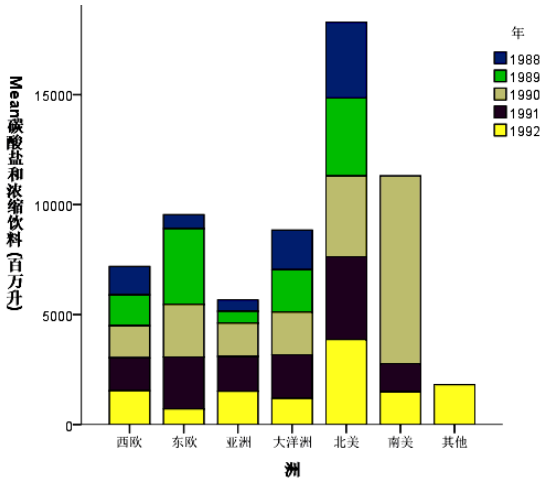


图 20-11 各洲每年碳酸盐和浓缩饮料平均产量

20.2.5 3-D条形图

- 1. 读取 data20-06 数据。按 Graphs→Legacy Dialogs→3-D Bar 顺序单击鼠标，打开 3-D Bar Charts 条形图主对话框，见图 20-2。
  - 2. SPSS 系统自动默认 Y 轴为数值变量轴，X 轴和 Z 轴分别为分类变量轴。在 X 和 Z 轴上分别选择 Groups of cases 选项。
  - 3. 单击 Define 按钮，打开 Define 3-D Bars 对话框，见图 20-12。将变量 salary 送入 Variable 栏中作为 Y 轴变量。在 Bar Represent 下拉菜单中选择统计量 Mean of values。有些统计量需要单击 Set Parameters 按钮，打开设定参数对话框，设定参数。
- 选择 college 作为 X 轴变量送入 X Category Axis 框中，Z 轴选择 gender 变量。
- 单击 OK 按钮，不同性别不同时期毕业生的初始薪酬比图，见图 20-13。

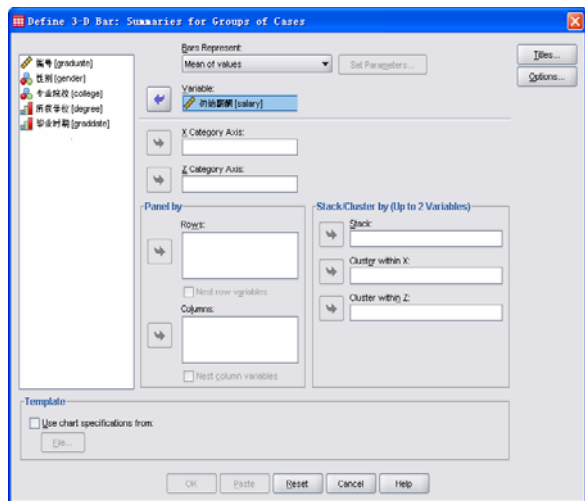


图 20-12 3-D 条形图对话框

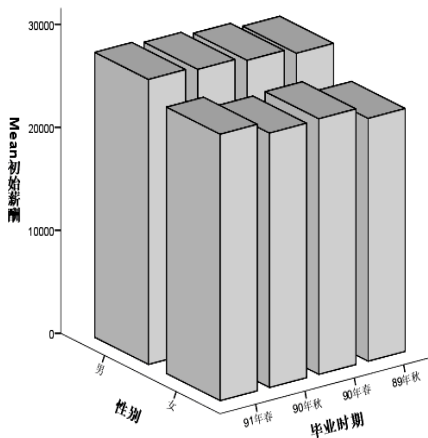


图 20-13 不同性别时期毕业生初始薪酬

## 20.3 线图、面积图和高低图

线图 (Line Charts), 又称曲线图, 是用线段的升降来说明现象变动情况的一种统计图, 它主要用于表示现象在时间上的变化趋势、现象的分配情况和两个现象之间的依存关系等。这里所指的线图均为纵横轴是算术刻度的普通线图。

面积图 (Area Charts) 是用线段下的阴影面积来强调现象变化的统计图。堆栈面积图可表示现象总体内部结构状况, 因此也称为结构曲线图。

高低图 (High-Low Charts) 是一种说明某些现象在单位时间内变化情况的统计图。它适合描述每小时、每天、每周等时间内不断波动的市场信息资料, 如股票、商品价格、货币牌价等, 高低图既说明某些现象在短时间内的变化, 也可说明它们长期的变化趋势。

### 20.3.1 选择图形类型

按 Graphs→Legacy Dialogs→line 顺序单击鼠标, 打开线图主对话框, 见图 20-14。

按 Graphs→Legacy Dialogs→Area 顺序单击鼠标, 打开面积图主对话框, 见图 20-15。

按 Graphs→Legacy Dialogs→High-Low 顺序单击鼠标, 打开高低图主对话框, 见图 20-16。在各对话框中选择图形的模式。

(1) 在如图 20-14 所示的线图对话框中选择线图模式

- Simple 单线图。用一条折线表示某种现象变动趋势的统计图。
- Multiple 多线图。用多条折线同时表示多种现象变动趋势的统计图。
- Drop-line 垂线图。反映某些现象在同一时期内差距的统计图。

(2) 在如图 20-15 所示的对话框中选择面积图模式



- **Simple** 简单面积图。用面积的变化表示某种现象变动趋势的统计图。
- **Stacked** 堆栈面积图。用不同种类的面积表示多种现象变动趋势和总体内部构成。

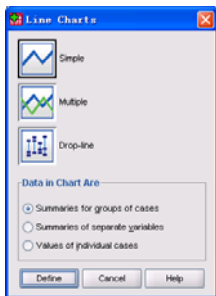


图 20-14 线图主对话框

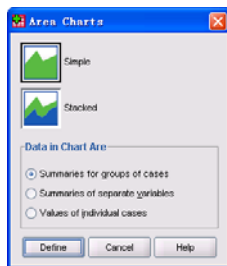


图 20-15 面积图主对话框



图 20-16 高低图主对话框

(3) 在如图 20-16 所示的对话框中选择高低图模式

- **Simple high-low-close** 简单高低收盘图，表示单位时间内某现象最高数值、最低数值和收盘数值。它适用于股票、期货等。它可说明每天最高、最低和收盘价。
- **Simple range bar** 简单极差图，或称为单式全距图，表明单位时间内某现象最高数值和最低数值。它与单式高低收盘图的区别是省去了收盘数值。
- **Clustered high-low-close** 分组高低收盘图，表示在单位时间内两个或两个以上现象的最高数值、最低数值和收盘数值。
- **Clustered ranger bar** 分组极差图，或称为复式全距图，它表示在单位时间内两个或两个以上现象的最高数值和最低数值。
- **Difference Line** 距限曲线图，它是说明两个现象在同一时间内相互变化对比关系的线性统计图。

统计量描述模式参见 20.2.1 小节。

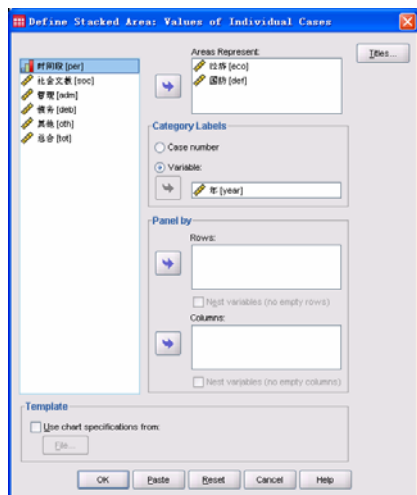
### 20.3.2 观测值模式堆栈面积图

在面积图主对话框中选择 **Stacked** 和 **Variables of individual cases** 选项，单击 **Define** 按钮，展开 **Define Stacked Area: Variables of Individual Cases** 观测值模式堆栈面积图对话框，见图 20-17(a)。本小节数据来自 data20-07 数据文件，例图为 1950—1985 年我国每年国防支出总和和经济建设支出面积图，见图 20-17(b)。主要操作步骤是：

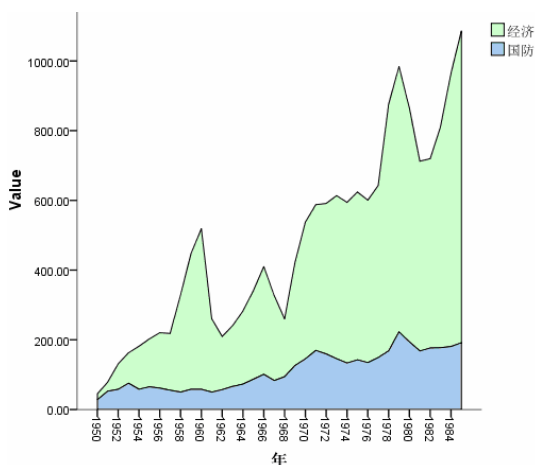
1. **Area Represent** 面积图表达统计量：选 **eco**、**def** 变量送入此框，描述该变量。
2. **Category Labels** 分类轴的标记和排列方式：

(1) **Case number** 以当前数据窗中的观测量序号为标记排列 **Area Represent** 框内变量的变量值，分类轴上变量值用阿拉伯数字标记。

(2) **Variable** 以某变量的变量值为标记排列 **Area Represent** 框内变量的变量值。在 **Variable** 框内选入 **year** 变量。



(a)

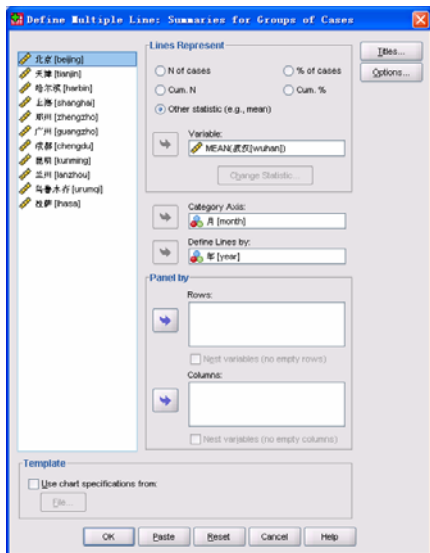


(b)

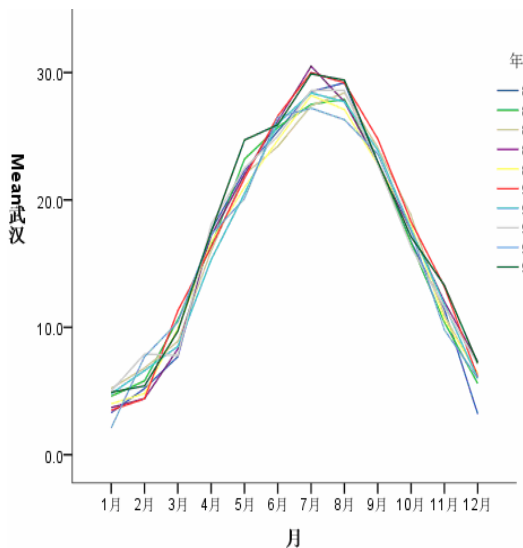
图 20-17 观测值模式堆栈面积图对话框及例图

### 20.3.3 观测量分类模式多线图

在线图主对话框中选择 **Multiple** 和 **Summaries for Groups of Cases** 选项，单击 **Define** 按钮，展开观测量分类模式多线图对话框，见图 20-18(a)。使用 data20-01 文件中的数据，例图为 1985—1994 年武汉月平均气温变化图，见图 20-18(b)。主要操作步骤是：



(a)



(b)

图 20-18 观测量分类模式多线图对话框及例图

- (1) Lines Represent 栏中选择 Other summary function, 选 wuhan 变量送入 Variable 栏。
- (2) Category Axis 分类轴变量框选择 month 变量, 具体操作见 20.2.2 小节。
- (3) Define Lines by 组变量: 在变量表中选择 year 变量送入 Define Clusters by 框中。

## 20.3.4 变量模式垂线图

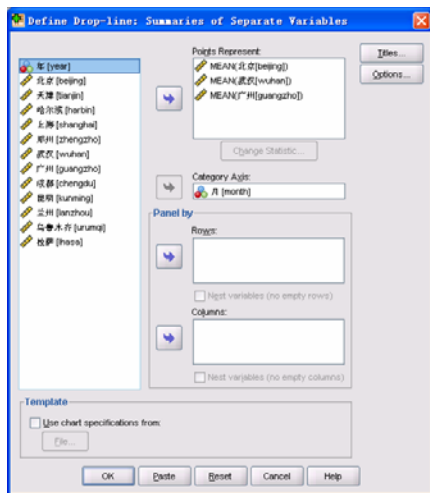
在线图主对话框中选择 Drop-line 和 Summaries of separate variables, 单击 Define 按钮, 展开 Define Drop-line: Summaries of Separate Variables 变量模式垂线图对话框, 见图 20-19(a)。例题数据为 data20-01, 例图为 1985—1994 年广州、北京、武汉月平均气温对比图, 见图 20-19(b)。主要操作步骤是:

- (1) 将变量 beijing、guangzhou、wuhan 送入 Points Represent 框中。
- (2) 将变量 month 作为分类轴变量送入 Category Axis 框中。单击 OK 按钮。

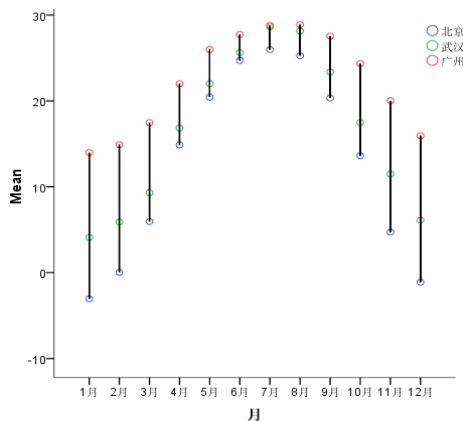
## 20.3.5 观测量分类模式简单高低收盘图

在高低图主对话框中选择 Simple high-low-close 和 Summaries for groups of cases 选项, 单击 Define 按钮, 展开观测量分类模式简单高低收盘图对话框, 见图 20-20(a)。例题数据为 data20-08, 例图 20-20(b)为 1996 年 4 月 1 日至 19 日地产类股票每天最高价、最低价和收盘价变化图。简单操作步骤是:

1. Bars Represent 高低收盘图表达统计量栏内选择 Other summary function 项, 再将 value 变量选入 Variable 框, value 变量的统计量为 MEAN, 改变统计函数参见 20.2.2 节。
2. Category Axis 分类轴变量: 选择 date 变量作为分类轴变量。
3. Define High-Low-Close by 确定高低收盘变量: 选择 hlc 变量作为高低收盘变量, 所生成条图的上端代表最高价, 下端代表最低价, 中间的方块代表收盘价。

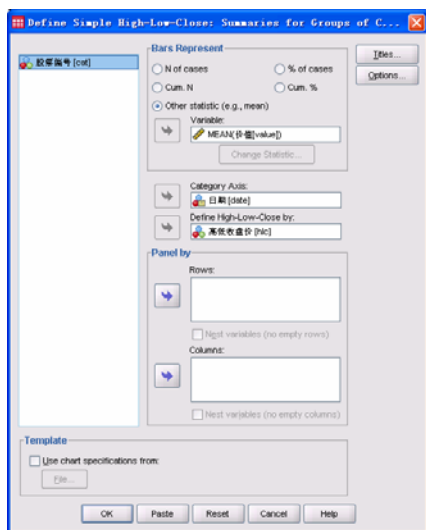


(a)

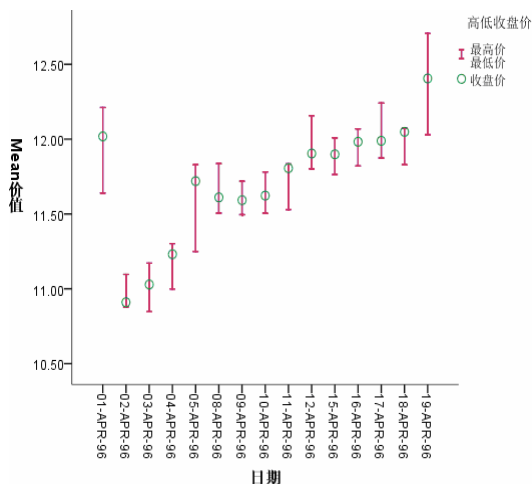


(b)

图 20-19 变量模式垂线图对话框及例图



(a)

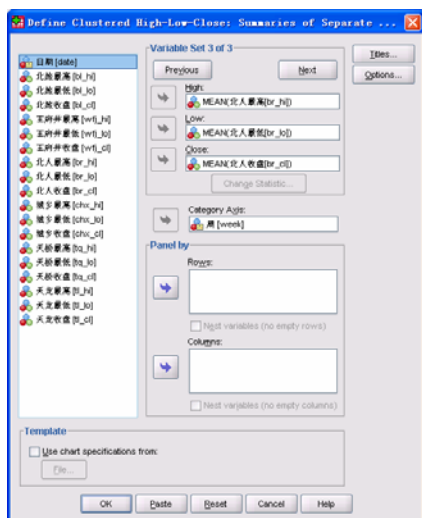


(b)

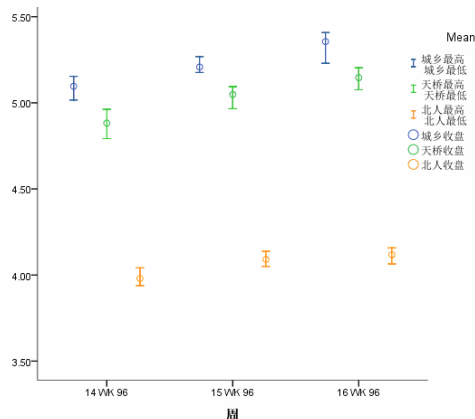
图 20-20 观测量分类模式简单高低收盘图对话框及例图

### 20.3.6 变量模式分组高低收盘图

在高低图主对话框中选择 **Clustered high-low-close** 和 **Summaries of separate variables** 选项，单击 **Define** 按钮，展开变量模式分组高低收盘图对话框，见图 20-21(a)。例题数据为 data20-09，例图 20-21(b)为 1996 年第 14、15、16 周城乡股票、北人股票以及天桥股票对比变化图。主要操作步骤是：



(a)



(b)

图 20-21 变量模式分组高低收盘图对话框及例图

1. High 框中的变量将作为条图的上端值; Low 框中的变量将作为条图的下端值。
2. Close 框中的变量将作为条图的方块, 是收盘值。
3. Category Axis 框内的变量 week 作为分类轴。

High 框和 Low 框中必须选有变量, 而 Close 框则可选或不选入变量, 如果在 Close 框内没有选入变量, 最后生成的图形就没有最后数值的标记(方块)。

Variable Set M of N within Clusters 显示  $N$  套变量组中的第  $M$  套变量: 当选择完一套变量后, 即在 High、Low、Close 框中分别选入了一套变量的最高价、最低价或收盘价变量后, 单击 Next 按钮并出现提示录入下一套变量。本例中录入三套变量(chx-hi、chx-lo、chx-cl, tq-hi、tq-lo、tq-cl 和 br-hi、br-lo、br-cl), 录入完第一套 chx 变量, 单击 Next 按钮并出现 Variable Set 1 of 1 within Clusters 提示; 如果录入完这三套变量, 文字提示将显示 Variable Set 3 of 3 within Clusters, 其含义为当前的这些变量是三套分组变量中的第三套变量。要修改第二套变量, 单击 Previous 按钮, 文字提示显示 Variable Set 2 of 3 within Clusters, 即为三套变量组中的第二套变量, 同时在相应的变量框内显示第二套变量的 High、Low 和 Close 变量。

### 20.3.7 观测量分类模式简单极差图

在高低图主对话框中选择 Simple Range Bar 和 Summaries for groups of cases 项, 单击 Define 按钮, 展开观测量分类模式简单极差图对话框, 见图 20-22(a)。例题数据为 data20-10, 例图 20-22(b)为 1996 年 4 月 1 日至 19 日工业股票和商业股票每日收市平均价对比图。主要操作步骤是:

1. Bars Represent 极差图, 表达统计量栏内选择 Other summary function 项, 再将 close 变量选入 Variable 框, close 变量的统计量为 MEAN, 改变统计函数参见 20.2.2 节。
2. Category Axis 分类轴框内选入 date 作为分类轴变量。
3. Define 2 Groups by 确定极差图两端变量。极差图两端各代表不同的变量值, 因此这个变量只能有两个变量值, 通过极差图的长短表示这个变量值的差距。选择 group 为两端变量, 在 Define 2 Groups by 框中显示变量名 group。

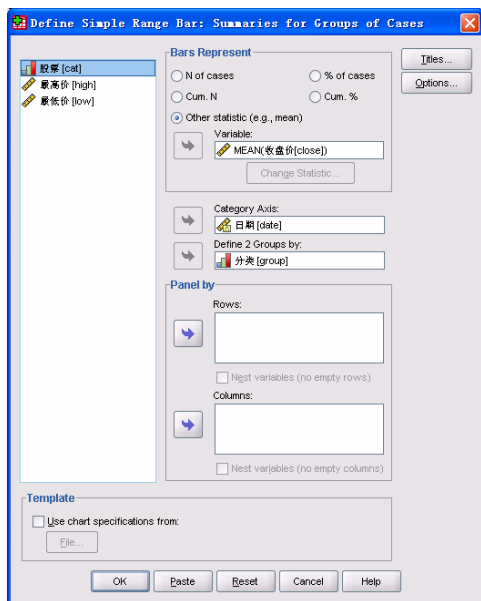
### 20.3.8 变量模式简单极差图

在高低图主对话框中选择 Simple Range 和 Summaries of separate variables 项, 单击 Define 按钮, 展开变量模式简单极差图对话框, 见图 20-23(a)。例题数据是 data20-09, 例图 20-23(b)为 1996 年第 14、第 15 和第 16 周天桥股票最高和最低价格对比图。主要操作步骤是:

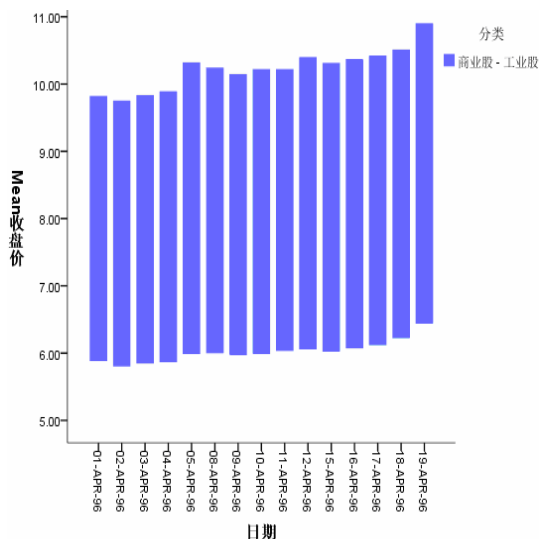
1. Bar Pair Represent 极差图表达的统计量: 极差图两端各表达两个不同变量。如果改变这两个变量的统计函数参见 20.2.2 节。

1st: tq-hi 变量作为第一被描述的变量; 2nd: tq-lo 变量作为第二被描述的变量。

2. Category Axis 分类轴变量: 选择 week 分类轴变量。

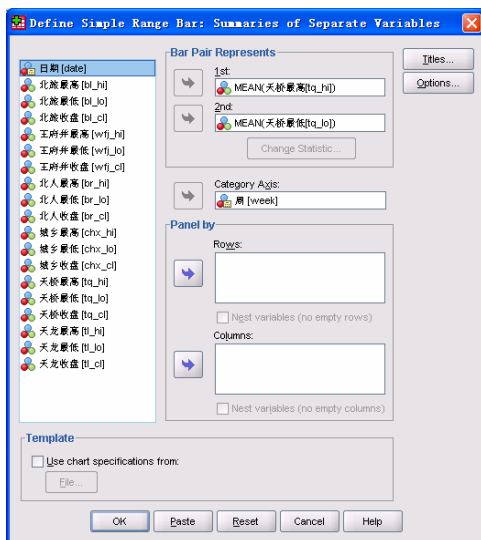


(a)

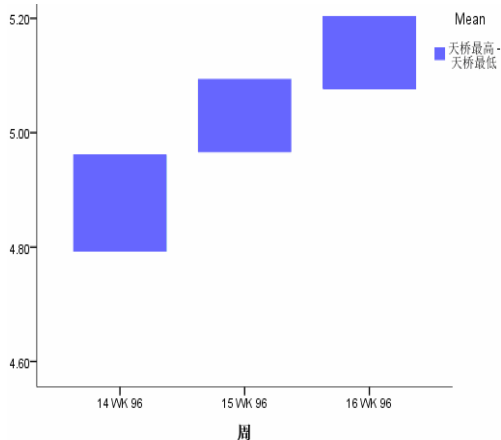


(b)

图20-22 观测测量分类模式简单极差图对话框及例图



(a)



(b)

图 20-23 变量模式简单极差图对话框及例图

20.3.9 观测值分类分组极差图

在高低图主对话框中选择 **Clustered Range Bar** 和 **Variables of individual cases**，单击 **Define**，展开观测值模式分组极差图对话框，见图 20-24(a)。例题数据 data20-09。例图 20-24(b)为 1996 年 4 月 1 日至 19 日每天北旅股票、北人股票和城乡股票最高价格与最低价格对比变化图。操作步骤重点为：

- 1. 在变量栏中选择 **bl\_hi** 变量进入 1st 框，选择 **bl\_cl** 变量进入 2nd 框。
- 2. 在变量栏中选择 **chxl\_hi** 变量进入 1st 框，选择 **chx\_cl** 变量进入 2nd 框。
- 3. 将 **br\_hi** 和 **br-cl** 变量分别进入 1st 和 2nd 框中，每步单击 **Next** 按钮进入下一步。
- 4. 在 **Category Labels** 栏中，选 **Variable** 项，将 **date** 变量送入其下的框中。

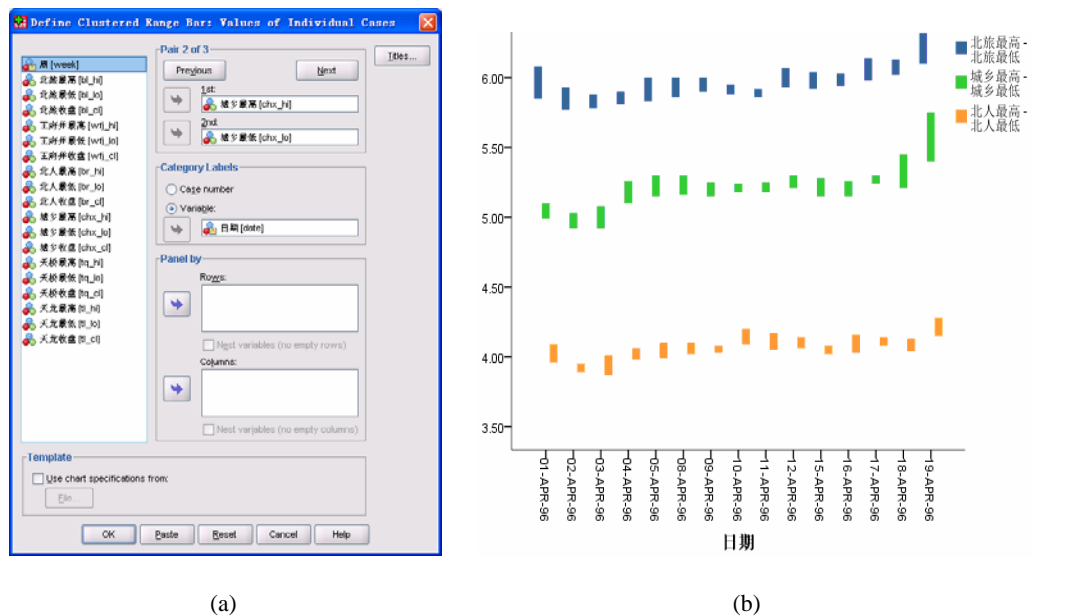
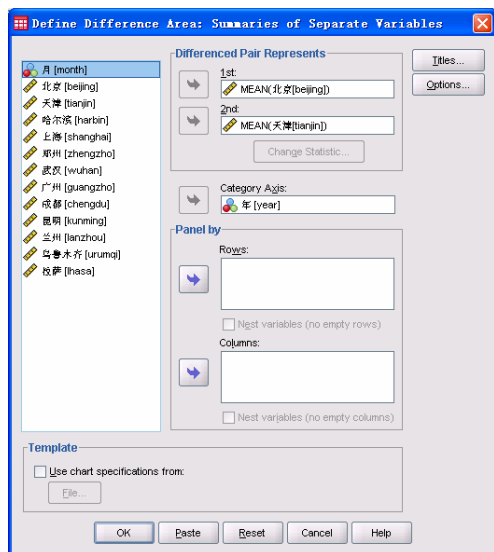


图 20-24 观测值模式分组极差图对话框及例图

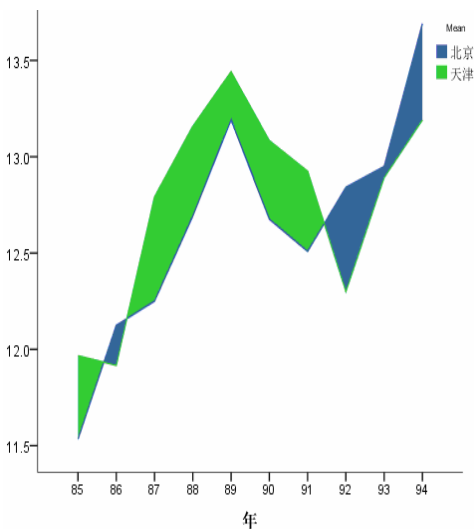
20.3.10 变量模式差分面积图

在高低图主对话框中选择 **Difference Area** 和 **Summaries of Separate Variables** 选项，单击 **Define** 按钮，展开定义变量模式差分面积图对话框，见图 20-25(a)。例题数据 data20-01。例图 20-25(b)为 1985—1994 年北京和天津年平均气温对比图。在图形中，浅色代表天津年平均气温，深色代表北京年平均气温，浅色在上表示天津年平均气温高于北京年平均气温，而深色在上表示北京年平均气温高于天津年平均气温。

主要操作步骤是：将 **Beijing** 和 **Tianjin** 变量分别选入 1st 和 2nd 框中，统计函数为 **Mean**。将 **year** 变量选入 **Category Axis** 框中。单击 **OK** 按钮。



(a)



(b)

图 20-25 变量模式差分面积图对话框及例图

## 20.4 圆 图

圆图 (Pie Charts) 又称饼图, 常用来表现构成比。以整个圆代表被研究现象的总体的, 按各构成部分占总体比重的大小把圆面积分割成若干扇形, 表示部分对总体的比例关系。

按 Graphs→Legacy Dialogs→Pie 顺序单击鼠标, 打开 Pie Charts 圆图主对话框, 见图 20-26。

由于 SPSS 系统在 Graph 菜单中只提供了单圆图, 所以仅有 3 种统计量描述模式。

### 20.4.1 观测量分类模式圆图

在圆图主对话框中选择 Summaries for Groups of Cases 项, 单击 Define 按钮, 展开 Define Pie: Summaries for Groups of Cases 观测量分类模式圆图对话框, 见图 20-27(a)。例题数据 data20-11。例图 20-27(b)为 1993 年俄罗斯每季度失业人口 (万人)。主要操作步骤是: Slices Represent 栏, 设置扇面表达的统计量。选择 Other summary function, 再选入 rus 变量进入 Variable 框, rus 变量的统计量为 SUM; Define Slices by 框, 确定扇面分类变量框内选入 sea 变量作为扇面分类变量。

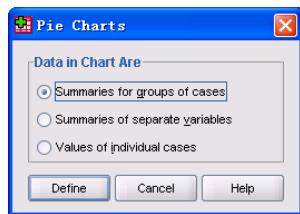
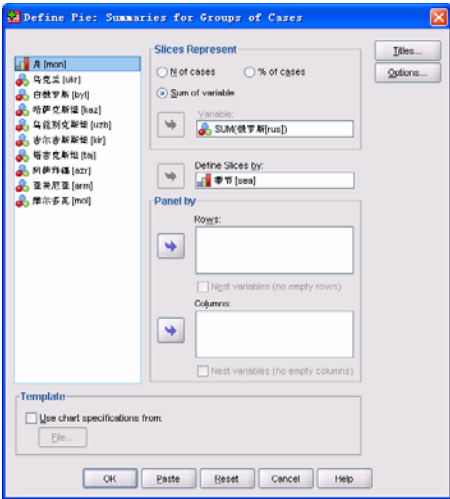
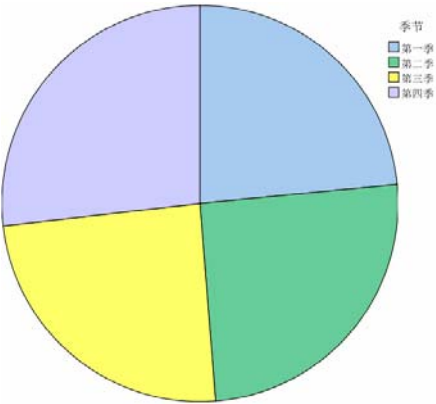


图 20-26 圆图主对话框





(a)

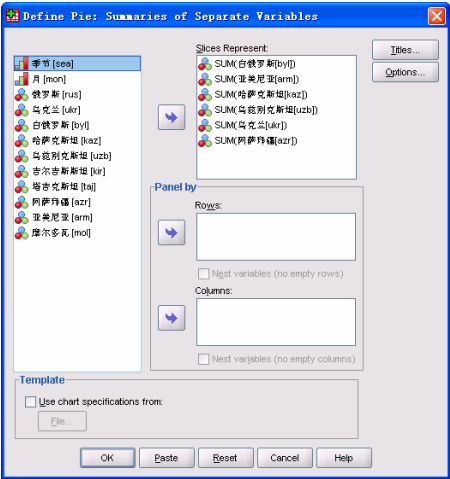


(b)

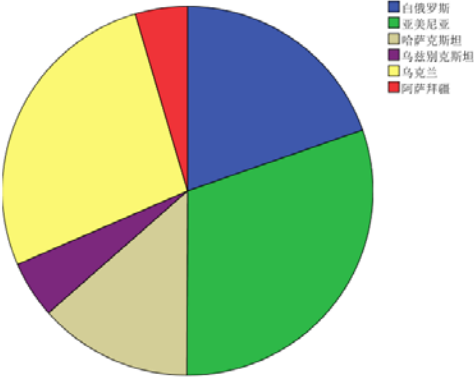
图 20-27 观测量分类模式圆图对话框及例图

20.4.2 变量模式圆图

在圆图主对话框中选择 Summaries of Separate Variables 项，单击 Define 按钮，展开 Define Pie: Summaries of Separate Variables 变量模式圆图对话框，见图 20-28(a)。数据 data20-11 数据文件，例图 20-28(b)为 1993 年部分独联体国家失业人口。



(a)



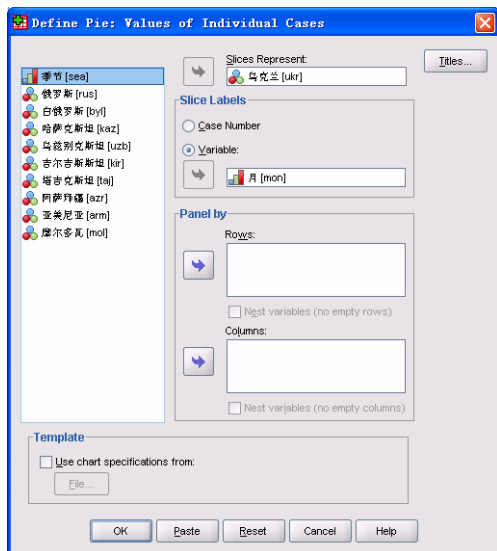
(b)

图 20-28 变量模式圆图对话框及例图

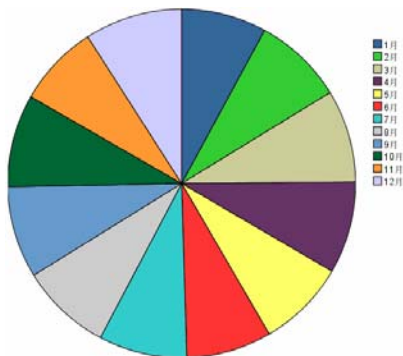
20.4.3 观测值模式圆图

在圆图主对话框中选择 Variables of Individual Cases 项，单击 Define 按钮，展开 Define

Pie: Variables of Individual Cases 各观测值模式圆图对话框, 见图 20-29(a)。例题数据 data20-11 数据文件。例图 20-29(b)为 1993 年乌克兰每月失业人口。



(a)



(b)

图 20-29 观测值模式圆图对话框及例图

## 20.5 箱图和误差条图

箱图 (Boxplots) 又称箱线图, 是一种描述数据分布的统计图形, 利用它可以从视觉的角度观察变量值的分布情况。箱图主要表示变量值的中位数、第 25 百分位数、第 75 百分位数等统计量, 其具体表示的统计量参见前面的章节。箱图可以从 Explore 统计过程中获得, 但是本节介绍的方法能够制作更复杂的箱图。

误差条图 (Error Bar Charts) 是一种描述数据总体离散的统计图形, 利用它可以从视觉的角度观察样本的离散程度, 误差条图表达平均数的置信区间、标准差或标准误。在误差条图中, 小方块表示平均数, 图形的两端为置信区间、标准差或标准误。

### 20.5.1 选择箱图和误差条图类型

通过箱图主对话框指定图形的类型, 按 Graphs→Legacy Dialogs→Boxplot 顺序单击鼠标, 打开 Boxplot Charts 箱图主对话框, 见图 20-30。

通过误差条图主对话框指定图形的类型, 按 Graphs→Legacy Dialogs→Error Bar 顺序单击鼠标, 打开 Error Bar 误差条图主对话框, 见图 20-31。

箱图、误差条图图式和统计量描述模式, 请参看 20.2.1 节。根据箱图、误差条图图式和统计量描述模式的选择组合, 共可生成 4 种不同类型的箱图和误差条图。

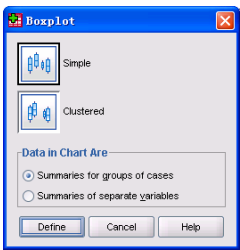


图 20-30 箱图主对话框

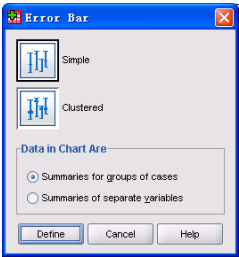
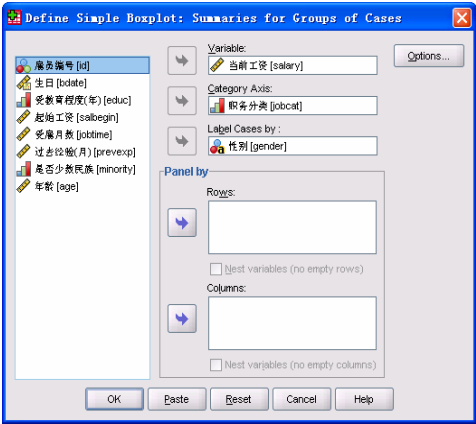


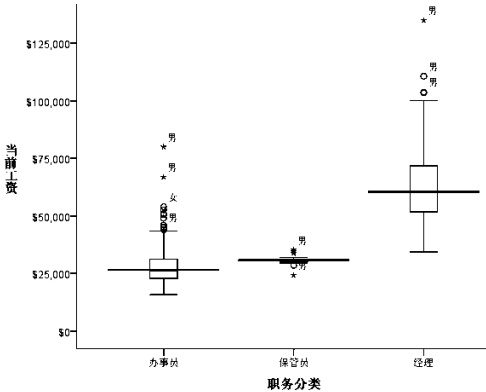
图 20-31 误差条图主对话框

20.5.2 观测量分类模式简单箱图

在箱图主对话框中选择 Simple 和 Summaries for groups of cases 项,单击 Define 按钮,展开定义观测量分类模式简单箱图对话框,见图 20-32(a)。例题数据 data20-12。例图 20-32(b)为不同岗位银行职员当前工资的箱线图。主要操作步骤是:选择要描述的变量 salnow 送入 Variable 框;选择分类轴变量 jobcat 送入 Category Axis 框;选择标识观测量的变量 sex 送入 Label Cases by 框。该变量值将对箱体外的观测量进行标识。“Male”标识男性,“Female”标识女性,见图 20-32(b)。



(a)

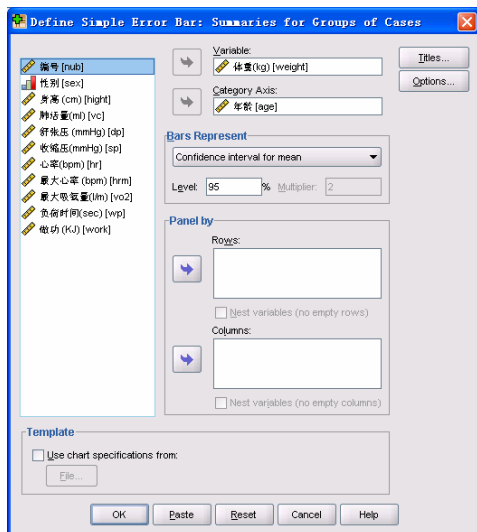


(b)

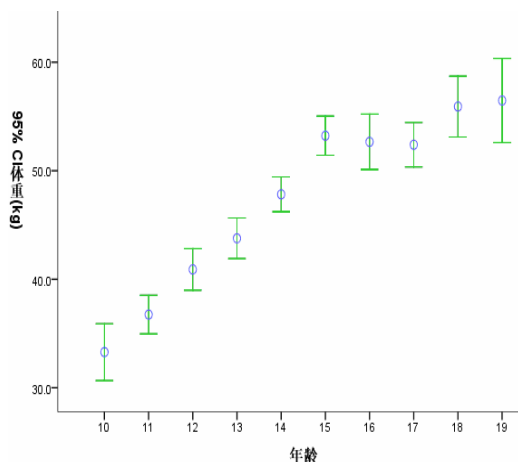
图 20-32 观测量分类模式简单箱图对话框及例图

20.5.3 观测量分类模式简单误差条图

在误差条图主对话框中选择 Simple 和 Summaries for groups of cases 观测量分类模式简单误差条图,单击 Define 按钮,展开相应的对话框,见图 20-33(a)。使用 data20-05 的数据。例图 20-33(b)为各年龄组受试者体重均值 95%置信区间的误差条图,在分类轴上 N 行的数值为每类的数量。主要操作步骤是:选择 weight 作为被描述变量送入 Variable 框中;选择 age 作为分类轴变量送入 Category Axis 框。



(a)



(b)

图 20-33 观测量分类模式简单误差条图对话框及例图

在 Bars Represent 条图表达统计量参数框中有 3 个选项：

- (1) Confidence interval for mean 均值置信区间，在 Level:\_% 框中输入需要的水平值。
- (2) Standard error of mean 均值标准误，Multiplier 框中输入均值标准误的倍数。
- (3) Standard deviation 标准差，Multiplier 框中可根据需要输入标准差的倍数。

## 20.5.4 变量模式简单箱图

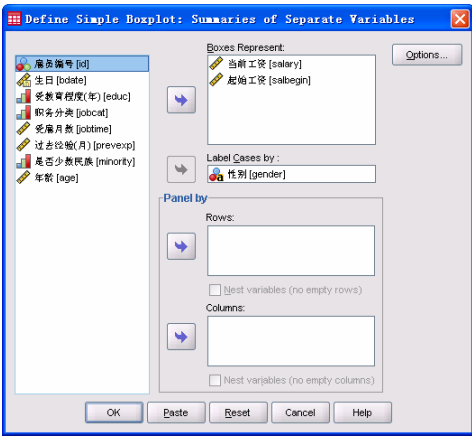
在箱图主对话框中选择 Simple 和 Summaries of separate variables 项，单击 Define 按钮，展开定义变量模式简单箱图对话框，见图 20-34(a)。例题数据 data20-12。例图 20-34(b) 为银行职员初始工资和当前工资的箱线图。

主要操作步骤是：选择 salbeg 和 salnow 变量送入 Boxes Represent 框作为要描述的变量；选择 gender 送入 Label Cases by 框，作为标识观测量的变量。

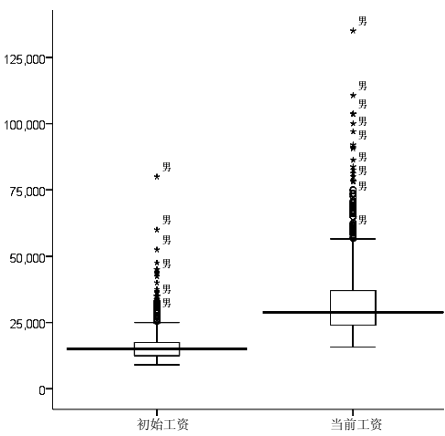
## 20.5.5 观测量分类模式分组误差条图

在误差条图主对话框中选择 Clustered 和 Summaries for groups of cases 项，单击 Define 按钮，展开定义观测量分类模式分组误差条图对话框，见图 20-35(a)。例题数据 data20-05，例图 20-35(b) 为男女各年龄组身高两倍标准差范围的误差条图。

主要操作步骤是：选择身高 height 变量送入 Variable 框中作为要描述的变量；选择年龄 age 变量作为分类轴变量送入 Category Axis 框中；选择性别 sex 变量作为标识类别的变量送入 Define Clusters by 框中。单击 OK 按钮。

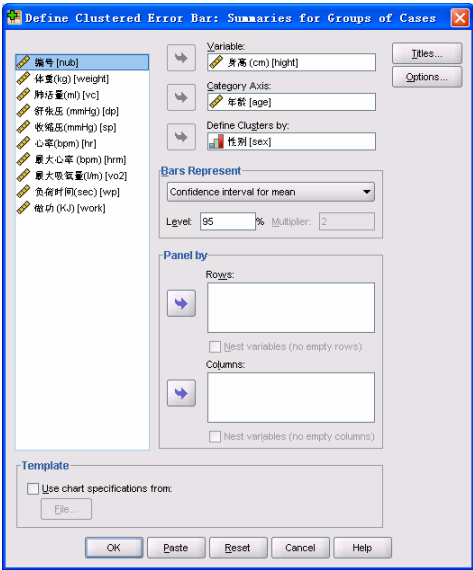


(a)

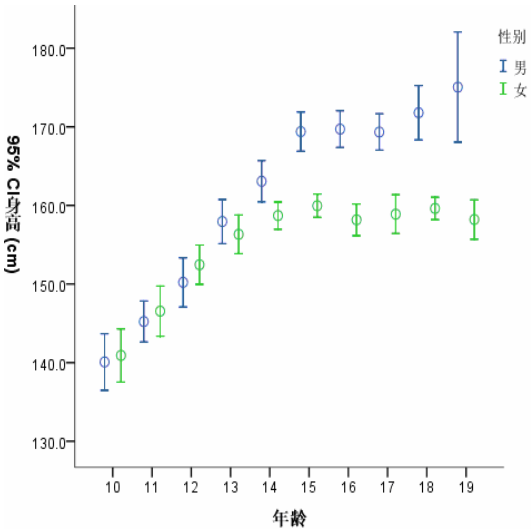


(b)

图 20-34 变量模式简单箱图对话框及例图



(a)



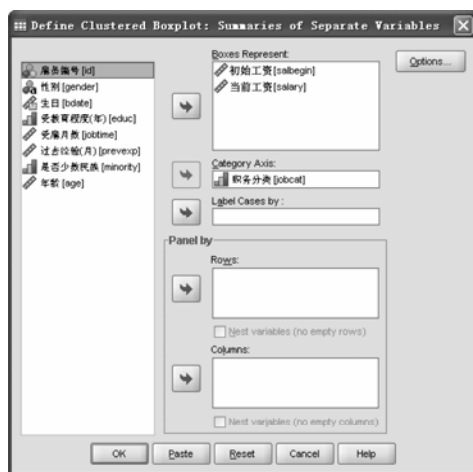
(b)

图 20-35 观测量分类模式分组误差条图对话框及例图

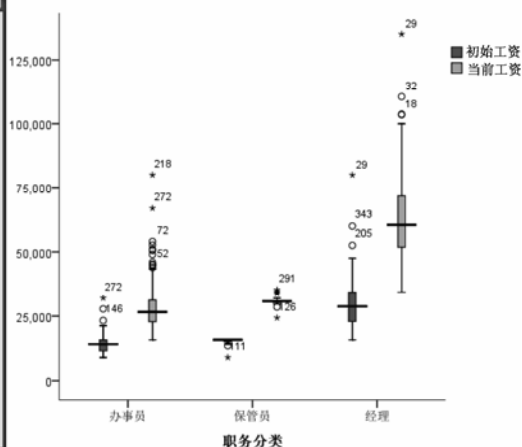
20.5.6 变量模式分组箱图

在箱图主对话框中选择 **Clustered** 和 **Summaries of separate variables** 项，单击 **Define** 按钮，展开定义多变量模式复式箱图对话框，见图 20-36(a)。数据文件 data20-12 中，分类轴变量选择 **jobcat**（职务等级），做箱图的变量选择 **salary**（当前工资）、**salbegin**（初始工资）。例图 20-36(b)为不同职务银行职员初始工资和当前工资的箱线图。

主要操作步骤是：选择初始工资 `salbegin` 和当前工资 `salary` 两个变量作为箱图要描述的变量送入 `BoxesRepresent` 框中；选择雇员职务 `jobcat` 变量作为分类轴变量送入 `Category Axes` 框中。单击 `OK` 按钮。



(a)



(b)

图 20-36 多变量模式分组箱图对话框及例图

## 20.6 散点图

散点图 (Scatterplots) 又称散布图或相关图, 是以点的分布反映变量间相关情况的图形, 根据图中的各点分布走向和密集程度, 大致可以判断变量之间协变关系的类型。

### 20.6.1 选择散点图图式

读者通过散点图主对话框指定散点图图式, 按 `Graphs`→`Legacy Dialogs`→`Scatterplot/Dot` 顺序单击鼠标左键, 打开 `Scatterplot/Dot` 散点图/垂点图主对话框, 见图 20-37, 共有 5 种散点图:

1. **Simple** 简单散点图, 显示一对相关变量的散点图。
2. **Overlay** 重叠散点图, 可显示多对相关变量的散点图。
3. **Matrix** 矩阵散点图, 在矩阵中显示多个相关变量之间的散点图。

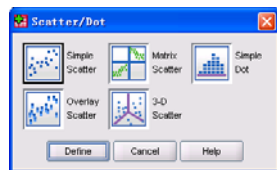
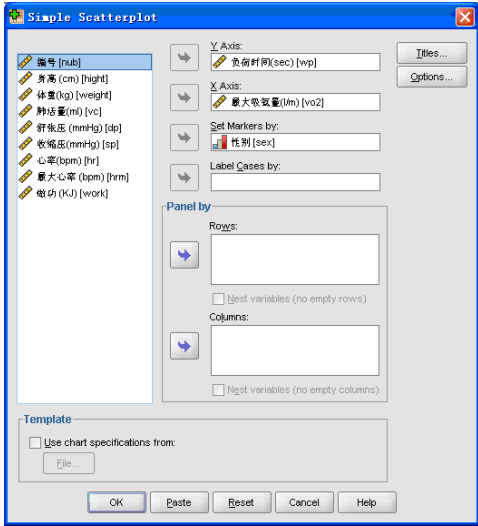


图 20-37 散点图主对话框

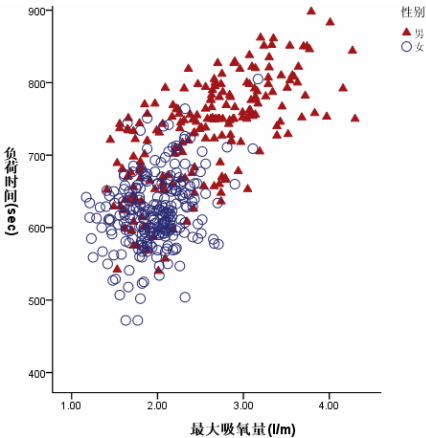
4. **3-D** 三维散点图, 显示三个相关变量之间的散点图。
5. **Simple Dot** 简单点图, 每个点代表一个观测量, 在图形中显示数值变量中各观测值在 *X* 轴上分布的图形, 该图也可看作一种散点图。

20.6.2 简单散点图

在散点图主对话框中选择 Simple 项，单击 Define 按钮，展开 Simple Scatterplot 简单散点图对话框，见图 20-38(a)。例题数据 data20-05，例图 20-38(b)为男女受试者最大吸氧量与负荷时间的简单相关图。主要操作步骤是：选择 wp 作为 Y 轴变量送入 Y Axis 框中；选择 vo2 作为 X 轴变量送入 X Axis 框中。选择 sex 送入 Set Markers by 框中



(a)



(b)

图 20-38 简单散点图对话框及例图

选择 sex 变量作为散点标记的变量送入 Set Markers by 框中。即用不同颜色或不同符号表示不同变量值。图 20-39 中以圆圈标记 女性的点，以三角形标记男性的点。

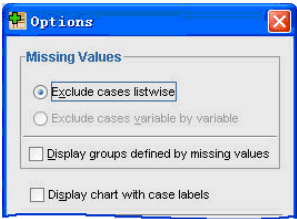



图 20-39 部分选择项对话框

还可以选择标识观测量点的变量送入 Label Cases by 中。每个点最多可用 20 个字符标识。还可以单击 Options 按钮，打开如图 20-39 所示的选择对话框，确定是否显示观测量的标识，选中 Display chart with case labels 所选择的标识变量才有效。为了使本小节的例图清晰，故均未选择该项。

20.6.3 重叠散点图

在散点图主对话框中选择 Overlay 选项，单击 Define 按钮，展开 Overlay Scatterplot 重叠散点图对话框，见图 20-40(a)。例题数据 data20-05。例图 20-40(b)为舒张压与体重、做功量与体重、身高与体重的重叠相关图。

在变量框中选择 Y-X 轴配对变量。第一个选择的为 Y 轴变量，第二个选择的为 X 轴

变量。送入 Y-X Pairs 框内。Variables 1 为第一个被选择的变量；Variables 2 为第二个被选择的变量，此显示记录了选择过程，方便了读者使用。然后再选择其他的变量对。本例选择了 dp-weight、work-weight 和 height-weight 三个变量对。如果想要调换 Y-X 轴变量的位置，则先选择变量对，再单击  置换 Y-X 按钮。

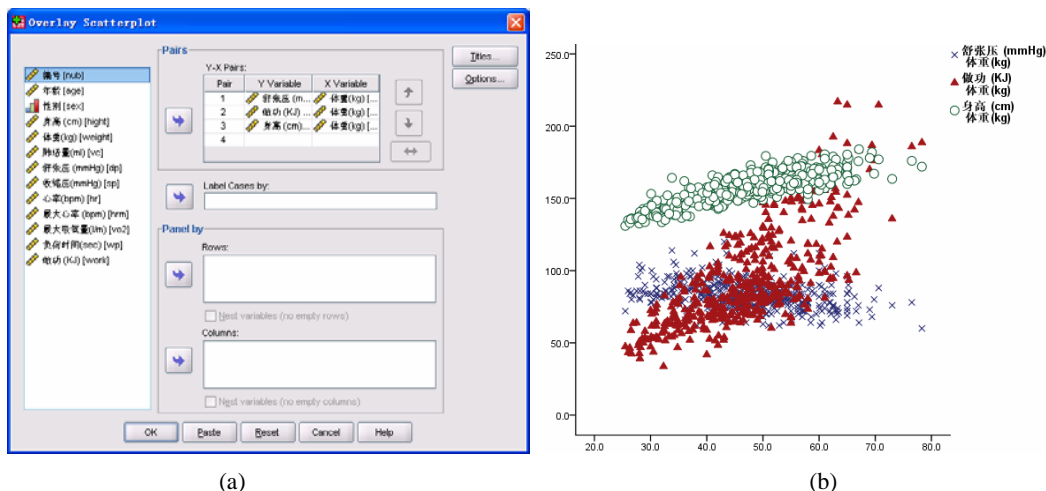


图 20-40 重叠散点图对话框及例图

## 20.6.4 矩阵散点图

在散点图主对话框中选择 Matrix 选项，单击 Define 按钮，展开 Scatterplot Matrix 矩阵散点图对话框，见图 20-41(a)。例题数据 data20-05。例图 20-41(b)为男女受试者最大吸氧量、肺活量和最大心率矩阵散点图。

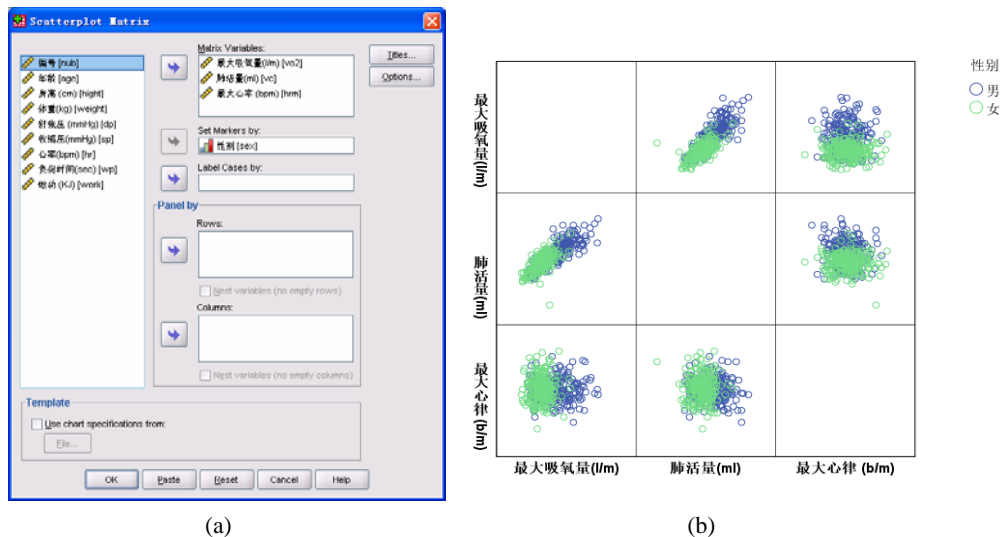


图 20-41 矩阵散点图对话框及例图



1. Matrix Variables 矩阵变量框内要选择两个或两个以上的变量，本例选择 vo2、vc 和 hrm 变量作为被描述变量。请读者注意矩阵变量框内的变量顺序与矩阵散点图对角线变量的顺序。

2. Set Markers by 设定散点标记，参见 20.8.2 小节。本例选择 sex 变量作为散点标记。

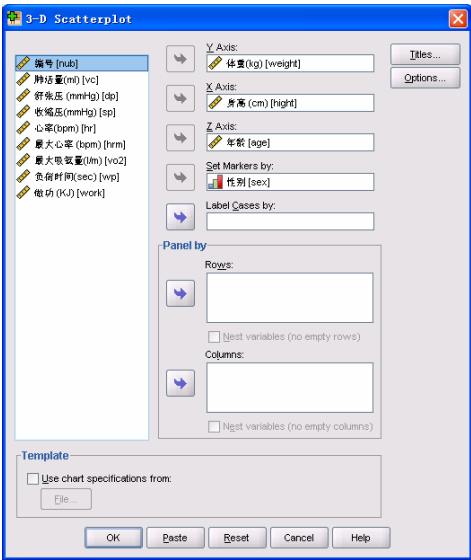
20.6.5 三维散点图

在散点图主对话框中选择 3-D 选项，单击 Define 按钮，展开 3-D Scatterplot 三维散点图对话框，见图 20-42(a)。例题数据 data20-05。例图 20-42(b)为男女受试者体重、身高和年龄三维散点图。

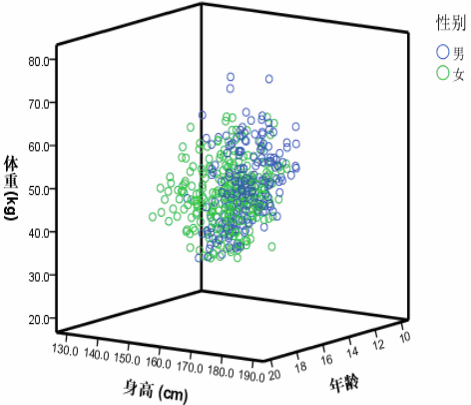
20.6.6 简单点图

在散点图主对话框中选择 Sample Dot 选项，单击 Define 按钮，展开 Define Simple Dot Plot 点图对话框，见图 20-43。例题数据 data20-05。主要操作步骤是：选择 vo2 变量作为被观测的变量送入 X-Axis Variable 框，该变量必须为数值型变量。通过点图在 X 轴上的堆栈情况，观察观测量的分布状态；单击 Options 按钮，打开选择对话框，见图 20-44，确定点图的分布形状：Asymmetric 非对称性堆栈分布，Symmetric 对称性堆栈分布，Flat 平行分布；选择 sex 变量，进入 Rows 框。

例图 20-45 为男女受试者最大吸氧量的非对称性堆栈分布点图。



(a)



(b)

图 20-42 三维散点图对话框及例图

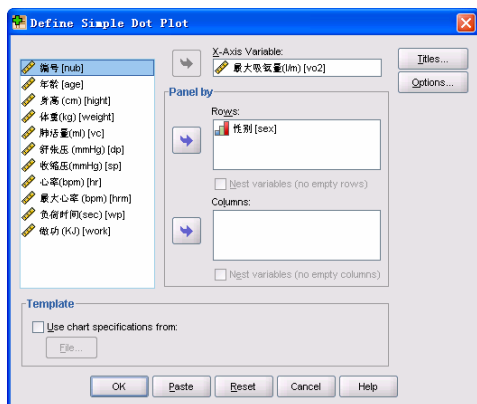


图 20-43 简单点图对话框

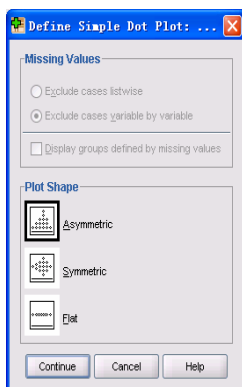


图 20-44 点图选择对话框

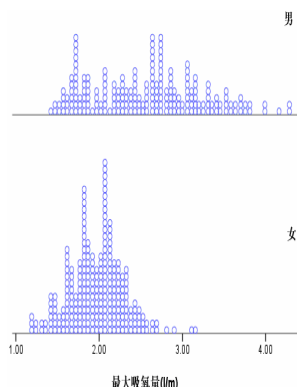


图 20-45 例图

## 20.7 直 方 图

直方图 (Histogram) 是以一组无间隔的直条, 表现频数分布特征的统计图, 直方图的每一条的高度代表相应组别的频数。

本节数据文件: data20-13 为某市 150 名 3 岁女童身高 (cm), 数据来源于《卫生统计》(周士楷, 人民卫生出版社)。data20-14 为 1971 年某市调查 200 例正常人血铅含量 ( $\mu\text{g}/100\text{g}$ ), 数据来源于《中国医学百科全书·医学统计学》(上海科学技术出版社)。

按 Graphs→Legacy Dialogs→Histogram 顺序单击鼠标, 打开 Histogram 直方图主对话框, 见图 20-46。

1. Variable 选择被描述的变量送入此栏。

(1) 本例使用 data20-13 数据, 选择变量 height 做描述变量。生成图 20-47(a), 为带有正态曲线的某市 150 名 3 岁女童身高直方图。

(2) 使用 data20-14 数据, 选择变量 pb 做描述变量。生成图 20-47(b)为带有正态曲线的某市 200 例正常人血铅含量直方图。

2. Display normal curve, 选择此项, 在生成的直方图上还显示正态曲线。本节两个例图都选择了此项。

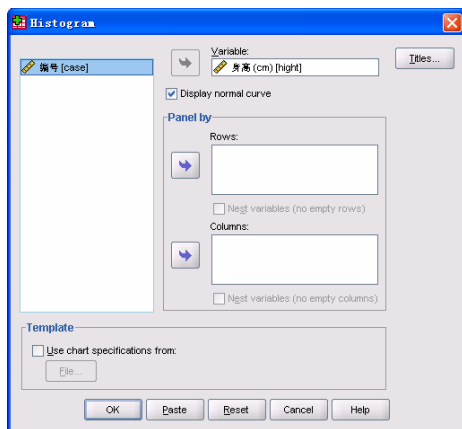


图 20-46 直方图对话框

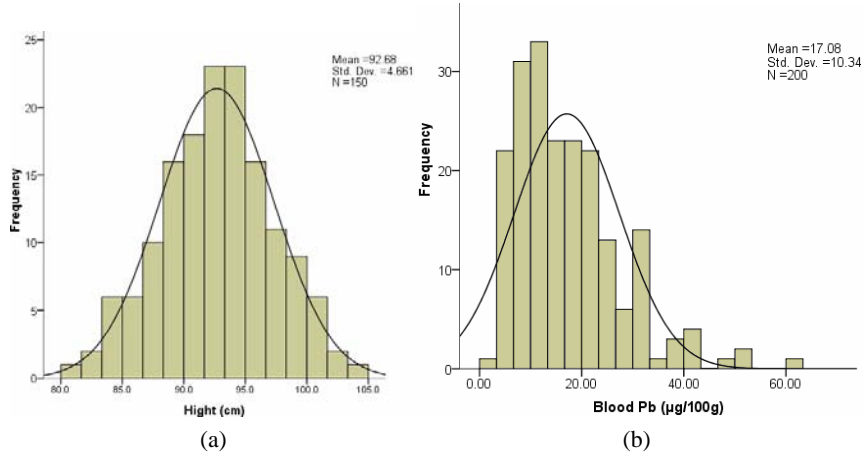


图 20-47 例图

## 20.8 交互图

### 20.8.1 交互式条形图、点图、线图和面积图

在 Graphs→Legacy Dialogs→Interactive 菜单下生成的图形为交互图形。图 20-48 为立体条形图，产生的效果往往与一般的二维统计图形不同。

按 Graphs→Interactive→Bar 顺序打开 Create Bar Chart 创建条图对话框。

按 Graphs→Interactive→Dot 顺序打开 Create Dots 创建点图对话框。点线图包括 Dot 点图、Line 线图、Ribbon 带图、Drop-Line 垂线图，这些图有共同特点，可以相互转换。

按 Graphs→Interactive→Area 顺序打开 Create Area Chart 面积图对话框。

#### 1. Assign Variables 指定变量选项卡

用鼠标单击 Assign Variables 选项卡，见图 20-49。在 Assign Variable 选项卡中有 3 种类型的图形。用鼠标单击右上角的图形类型按钮，展开图形类型选项：2-D Coordinate 二维坐标图、3-D Coordinate 三维坐标图和 3D Effect 三维效果图。

(1) 选择图形排列方式。当选择二维图形时，可选条、点、线图，有两种排列方式，分类轴是纵轴单击按钮 、；分类轴在横轴上单击 或 按钮。生成横排或纵排图。

(2) 选择变量进入轴框。从左侧的变量表中，将要被描述的变量分别选入坐标轴框内。一般情况下 X1 和 X2 轴框内为分类变量，Y 轴框内为尺度变量。展开对话框下部 Bar (Line、Dot) Represent 列表框，可以在其中选择 Y 代表的描述性统计量。

选中 Bar Represent 栏中的 Display Key 复选项，将在图形上显示图解说明。

(3) 尺度变量和分类变量均可以作为 Legend 图例变量，但它们表达的含义不同。尺度变量作为图例变量时，图例中变化的颜色或样式代表变量数值的范围。以这种

方式生成的条形图，实际上条的长度、颜色表达两种含义。

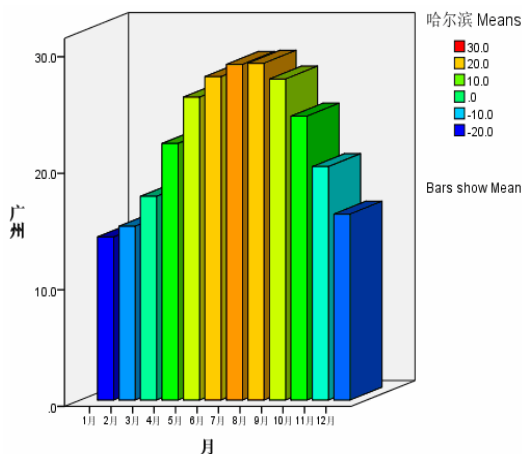


图 20-48 交互式条形图

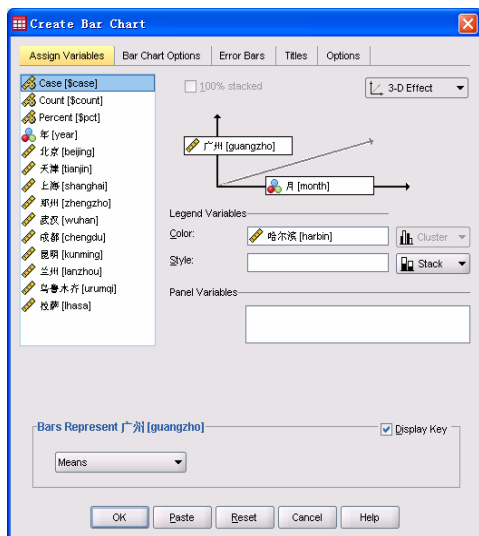


图 20-49 创建条形图对话框中指定变量选项卡

将分类变量选入 **Color** 颜色框内，图例中不同的颜色表示该 **Color** 变量的不同的分类；若将分类变量选入 **Style** 样式框内，图例中不同的样式表示 **Style** 变量的不同的分类。

(4) 选择条形图图式。在 **Color** 和 **Style** 框中选择了分类变量，就可以选择不同的条形图图式。如果在 **Color** 和 **Style** 框中选择了分类变量，将自动遮蔽条形图图式选择。

- **Cluster** 按钮，分组条形图，即复式条形图。
- **Stack** 按钮，分段条形图，即堆栈式条形图。
- **100% Stack** 复选项，百分比分段条形图，即条形图长度代表该分类的百分比构成。

(5) 选择 **Panel** 群组变量。进入 **Panel** 框中的变量，以其变量值作为分组的依据，生成多个图形，其目的是为了组群之间方便比较，读者可以为每个分类变量或分类变量与其他分类结合创建分离的群组图。

2. **Bars** 条块修饰选项卡，见图 20-50。只在 **Interictive** 中选了 **Bar**，才有此选项卡。

(1) **Bar Baseline** 条图基线栏。设立基线的目的是以基线为标准强调条图之间的差异，大于基线的条图立在基线上，小于基线的条图悬挂在基线下。

- ① **Automatic** SPSS 自动设置基线。
- ② **Custom** 在 **Custom** 框中输入数值，自定义基线的位置。

(2) **Bar Labels** 条图标签栏中，选中 **Count**，标注分类中的观测测量数；选中 **Value**，标注该分类变量中某个分类的统计量。

3. **Dots and Lines** 点图和线图选项卡，只在选择了三级菜单的 **Dot**、**Line** 才有点线选项卡，见图 20-51。

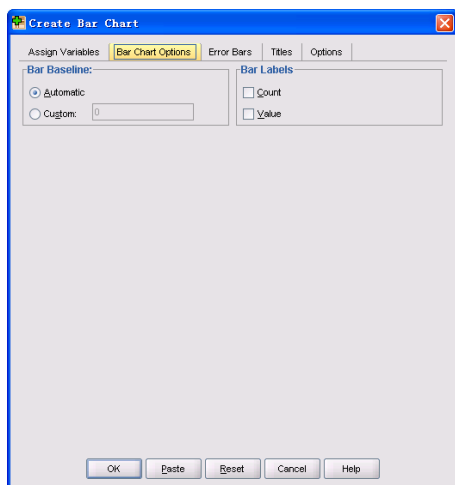


图 20-50 创建条形图对话框条形图修饰选项卡

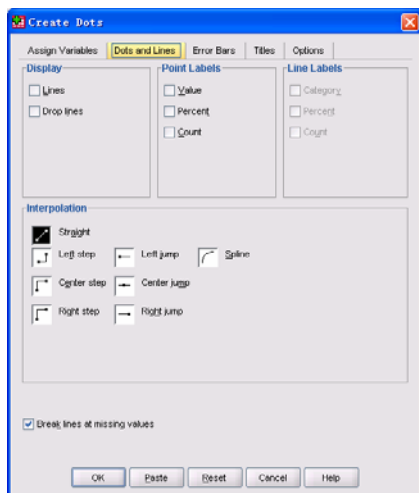


图 20-51 创建点图对话框中点线选项卡

(1) **Display** 显示点线栏中，选中 **Lines**，显示各点之间的连线。选中 **Drop lines**，在每个分类变量值点上显示各变量点的垂直连线。

(2) **Point Labels** 点图标签栏中，选中 **Value**，标注该点所代表的统计量。选中 **Percent**，标注该点所在分类占全部分类的百分比。选中 **Count**，标注该点所在分类中的观测量数。

(3) **Line Labels** 线图标签栏中，选中 **Category**，标注分类变量值。选中 **Percent**，标注分类变量中某个分类的百分比。选中 **Count**，标注某个分类的观测量数。

(4) **Interpolation** 添加连线栏。可选择不同的点间连线方式以及前三项所选值的标注位置。

(5) **Break lines at missing values** 分断线图图中的缺失值。

4. **Error Bars** 误差条图选项卡，见图 20-52。

(1) 选择 **Display Error Bars** 复选框后，在 **Confidence Interval** 栏中选择统计量，如可信区间、标准差和标准误。并可拖动游标确定参数，也可直接在参数框中输入参数。

(2) 在 **Shape** 栏中选择误差条图的形状和误差条图的帽宽。

5. **Titles** 图题注释选项卡，见图 20-53。

在 **Chart Title** 框中输入图的主标题内容，在 **Chart Subtitle** 框中输入图的子标题，在 **Caption** 框中输入图下注释。

6. **Options** 选择选项卡，见图 20-54。

在 **Options** 选项卡上可以选择条形图的预设格式和坐标轴的长短。

(1) **Chart Template**，图形模板框。可单击 **Browse** 按钮，打开浏览对话框查找所需要的图形模板文件。如果选择多个模板，可以通过  $\uparrow \downarrow$  按钮，将要使用的模板放置最顶层，也可将不需要的模板文件选中后，单击  $\times$  按钮删除。

(2) 在 **Chart Template** 框下 **Axis** 栏中，可对每个坐标轴的长度进行调整。

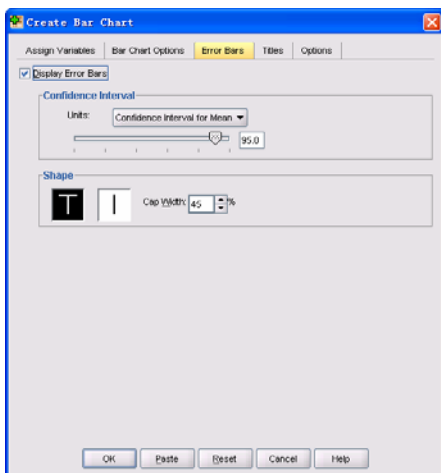


图 20-52 误差条图选项卡

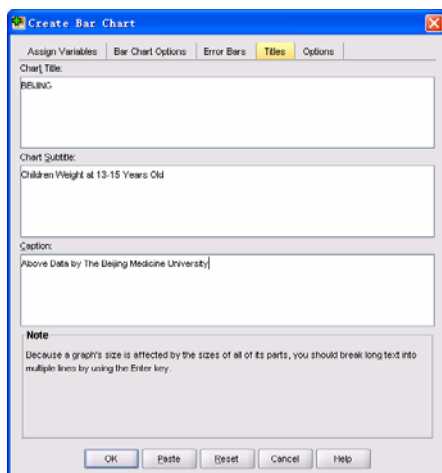


图 20-53 图题注释选项卡

(3) Categorical Order 分类次序栏，可以选择分类值在坐标轴上的排列次序。

① 在 Variable 下拉列表中显示所有的分类变量。选择一个排序分类变量。

② 在 Order by 框中确定排序的依据。

如果选择了 Values 值、Label 变量标签、Occurrence 出现的顺序、Counts 计数这 4 个特性之一，在 Variable 框中变量依所选特性进行排序；如果选择除上述 4 个选项以外的其他项，还可以在 Of 框选择数据文件中其他变量，即用其他变量的综合函数对分类值进行排列。Type 提示所使用变量的类型。

③ 选择了排列依据后，可在 Sort 栏中选择分类值排列的顺序，Ascending 升序、Descending 降序。选中 Exclude empty categories，将去掉没有数据的分类。

(4) 在 Scale Range 刻度范围栏中，选择坐标轴上的刻度范围。

① 在 Variable 框显示数据文件中所有的数值变量。选择刻度变量，确定刻度范围。

② 选择指定刻度范围的方式。默认为 Auto，自动确定刻度范围。如果不选 Auto 项，则需要在 Minimum 和 Maximum 参数框中分别指定最小和最大值。

## 20.8.2 交互式圆图

按 Graph→Interactive→Pie 展开三级菜单。有 3 个子命令，即 Simple 单圆图、Stacked

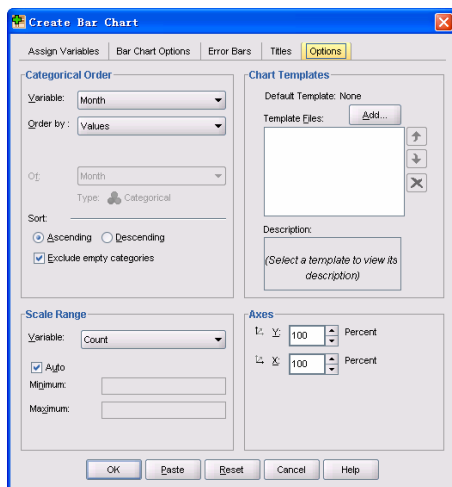


图 20-54 选择选项卡

堆栈圆图、Plotted 细分圆图。堆栈圆图是多个单圆图的组合，操作方法类似。

1. Stacked 堆栈圆图

(1) Assign Variable 指定变量选项卡，见图 20-55。

- ① 选择图形类型。从 2-D Coordinate 二维圆图和 3-D Effect 三维效果图中选择一项。
- ② 选择变量进入轴框。从变量列表中，将分类变量选入上面的 Slice by 框，并选择分类变量值以颜色 Color 区分或图案 Style 区分。将被描述的变量选入下面的 Slice 统计量变量框。

③ 选择另一个分类型变量进入 Stack by 堆栈变量框，作为圆图的堆栈分类。

④ 选择 Panel 群组变量。

(2) Pies 圆图修饰选项卡，见图 20-56。

① Slice Labels 扇形标签栏

- 显示标签内容，即选择各扇区标注的内容：Category 分类名称、Value 变量值、Count 计数和 Percent 百分比。
- Location 标签的位置，All Inside 在圆图之内，All Outside 在圆图之外，Text Inside, Number Outside 文字在圆图之内、数值在圆图之外，Number Inside, Text Outside 数值在圆图之内、文字在圆图之外。

② Cluster Labels 复式标签栏也分为标签内容和标签位置。标签内容的选择见 Slice Labels 的说明。Location 标签的位置选项有 Upper Right 圆图右上方、Upper Left 圆图左上方、Lower Right 圆图右下方、Lower Left 圆图左下方。

③ Position 扇形排列的方向及起始点的位置

- Directions 扇形排列的方向，分顺时针、逆时针方向排列。默认顺时针方向排列。
- Start 扇形的起始点的位置，从左至右为 12 点、3 点、6 点和 9 点。

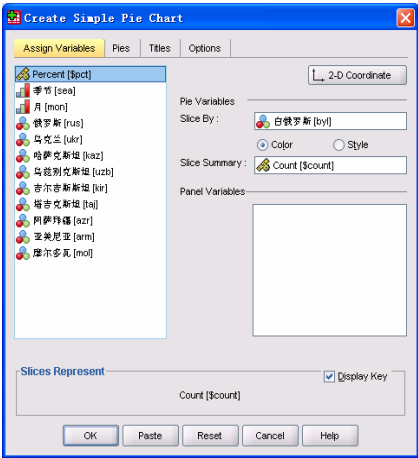


图 20-55 堆栈圆图对话框中指定变量选项卡

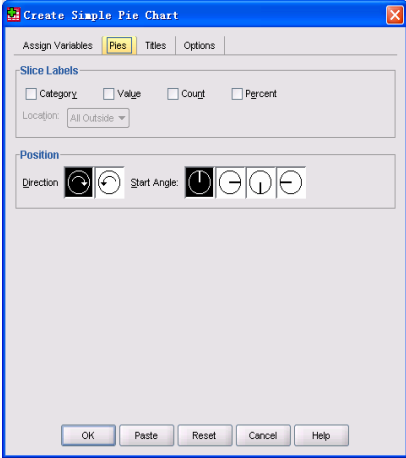


图 20-56 堆栈圆图对话框中圆图选项卡

## 2. Plotted 坐标式圆图

坐标式圆图是在一个坐标中一系列的圆图，见图 20-57，该图采用 data20-12 数据。坐标式圆图的操作除 Assign Variable 指定变量选项卡与复式圆图的功能不同，其他选项卡与复式圆图操作相同。

用鼠标单击 Assign Variable 选项卡，显示该选项卡的内容，见图 20-58。

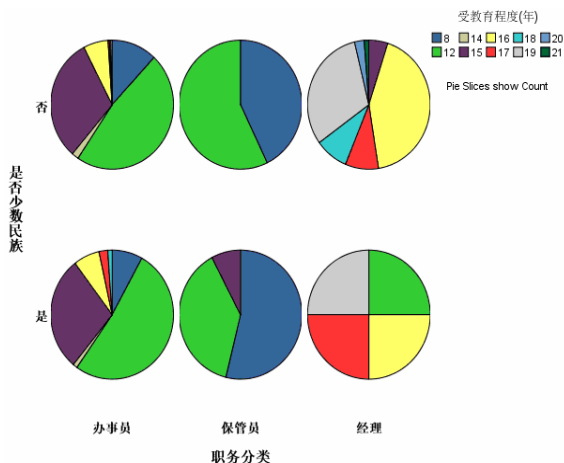


图 20-57 坐标式圆图

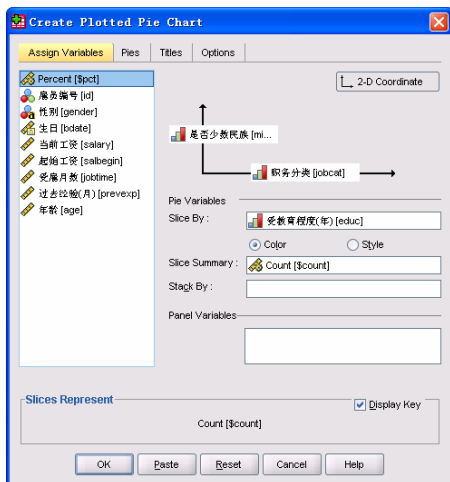


图 20-58 坐标式圆图对话框中指定变量选项卡

(1) 选择图形类型：2-D Coordinate 二维坐标式圆图、3-D Coordinate 三维坐标式圆图、3-D Effect 立体效果坐标式圆图。

(2) 上方的 Slice 框为分组变量框，选入分类型变量。

(3) 下方的 Slice 框为统计量框，在本框内可以选入尺度型变量，然后再在 Slice Represent 栏中选择要表达的统计量，系统默认 Count 统计量。

(4) X1 轴框为横轴细分组变量框，一般选入分类变量作为坐标轴上的横轴分类值。Y 轴框为纵轴细分组变量框，一般选入分类变量作为坐标轴上的纵轴分类值。

在上述两框内既可以选入分类型变量也可以选入尺度型变量，如果这两个框内选择的是尺度型变量，生成的图形近似于散点图。

(5) Cluster 框为复式变量框，作为圆图的复式分类。

### 20.8.3 交互式箱图和误差条图

按 Graphs→Interactive→Boxplot 顺序展开 Create Boxplot 创建箱图对话框。按 Graphs→Interactive→Error Bar 顺序展开 Create Error Bar Chart 创建误差条图。交互式箱图和误差条图的操作与交互式条图基本相同，此处只说明不同之处。

#### 1. Boxes 箱图修饰选项卡

单击创建箱图对话框中 Boxes 选项卡，显示该选项卡的内容，见图 20-59。



(1) Boxes Display 箱图展示栏中有 3 个复选项，要求显示：Outliers 奇异值、Extremes 极值、Median line 中线。

(2) Whisker Caps 触须线帽形栏有 2 个复选项：简洁帽形，无帽形。

(3) Box Base 箱图柱形栏有两个选项，只对三维图形有效，即 Square base 方柱状和 Circular 圆柱状。

(4) Display count labels 显示计数标记，选中此项，将在图形中显示每个分类的数量。

## 2. Error Bars 误差条形图修饰选项卡

误差条形图有 4 个选项卡，在 Assign Variables 选项卡上指定误差条形图代表误差的范围，如标准差、标准误、可信区间等。

单击创建误差条形图对话框中 Error Bars 选项卡，显示该选项卡的内容，见图 20-60。

(1) Bar Labels 误差条形图的标签栏中有两个复选项，即 Mean 均数和 Count 计数。

(2) Shape 误差条形图的形状栏，即 Plain Error Bar Caps 简洁帽形和 No Error Bar Caps 无帽形。

(3) Display symbol 显示均数标记，选中该项，将在图形中显示均数标记。

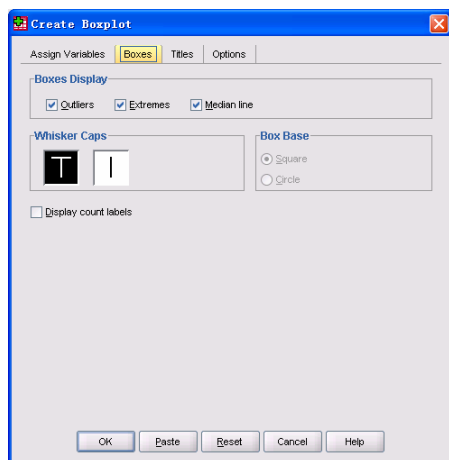


图 20-59 箱图对话框中的条图修饰选项卡

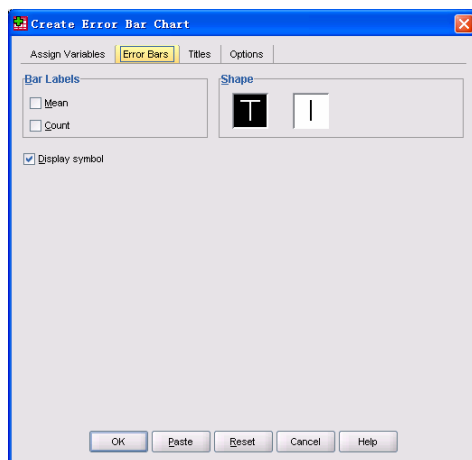


图 20-60 误差条图对话框中的条图修饰选项卡

## 20.8.4 交互式直方图

根据 data20-05 数据创建变量 VO2（最大吸氧量）的交互式累计直方图。重点步骤是按 Graphs→Legacy Dialogs→Interactive→Histogram 顺序展开 Create Histogram 创建直方图对话框，见图 20-61。将最大吸氧量变量 vo2 置于横轴，且要选择 cumulative histogram 项，生成图 20-62。在该对话框中共有 4 个选项卡，本节只说明 histogram 直方图选项卡。

1. Normal curve，选中该项，在生成的直方图中还生成一条正态曲线。

2. Set interval and start point for the，确定直方图宽度和起始点的坐标轴参数框，如果

是二维图形只能选择 X1 轴，如果是三维图形可以选择 X1 或 X2 轴。

3. **Interval Size** 图条宽度栏共有 3 个选项，如果选中 **Set interval size automatic** 项，图条尺寸自动设置。不选此项，激活 **Number of interval** 图条数量框、**Width of interval** 图条宽度参数框，读者可以根据需要调整图条的数量和宽度。最多可以生成 250 个图条。

4. **Start Point** 图条起始点栏，读者可以通过滑动游尺或直接在参数框内输入数值确定图条起始点，即从第一个图条宽度的百分比作为起始点。例如选择 30%，也就是从第一个图条宽度的 30% 处作为图条的起始点。

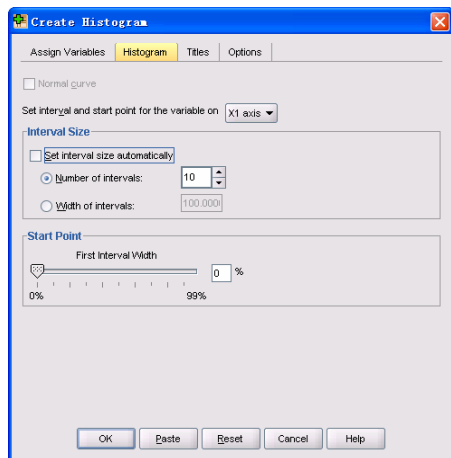


图 20-61 创建直方图对话框中直方图修饰选项卡

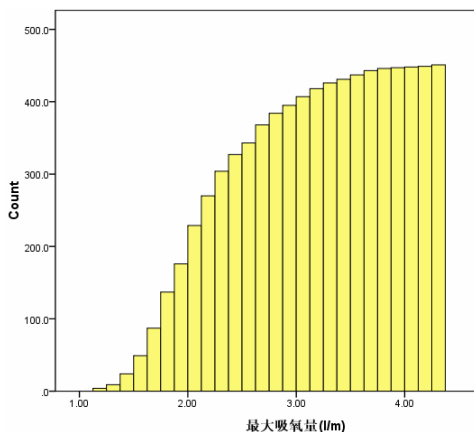


图 20-62 交互式累计直方图

## 20.8.5 交互式散点图

按 **Graphs**→**Legacy Dialogs**→**Interactive**→**Scatterplot** 顺序展开 **Create Scatterplot** 创建散点图对话框，该窗口共有 5 个选项卡，本节只说明 **Fit** 和 **Spikes** 选项卡。

### 1. **Fit** 拟合曲线选项卡

单击创建散点图对话框中的 **Fit** 选项卡，显示该选项卡的内容，见图 20-63。

**Method** 下拉菜单中选择拟合方法。选择 **None**，**Fit** 选项卡上的其他选项将被屏蔽。

(1) **Regression** 项选择用回归方法拟合，出现的 **Fit** 选项卡选项如图 20-63 所示。

① **Include constant in equation** 在图形中生成一个带常数项的回归方程。

② 在 **Prediction Lines** 预测线栏：若在 **Confidence Interval** 置信区间参数框输入确

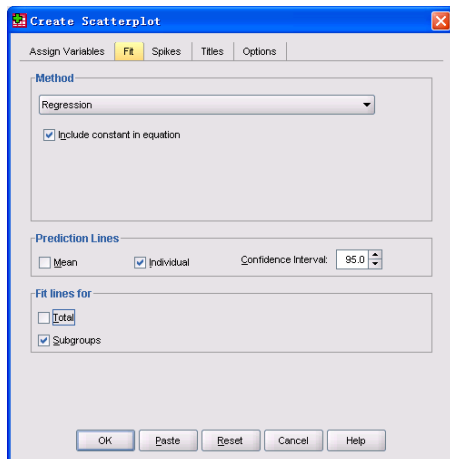


图 20-63 创建散点图对话框回归线选项卡

定的参数值，除生成回归线外：

- 选择 Mean 均线，生成均线的置信区间线。
- 选择 Individual 个点，还生成散点分布的置信区间线。
- ③ 在 Fit lines for 栏内有 Total 和 Subgroups 复选项。
- 选 Total，所有点生成一条回归线。
- 选 Subgroups，按细分组的散点分别生成回归线。

(2) 选择 Mean 项的 Fit 选项卡，它与 Regression 的 Fit 选项卡基本相同。

(3) 选择 Smoother 项出现的 Fit 选项卡选项，见图 20-64。

① Kernel 框，选择了 Smoother（LLR 局部线性回归）后，在该框内指定 kernel 项。

② 在 Bandwidth 带宽栏中的 X1 和 X2 轴上，指定拟合曲线带宽。

③ Use same bandwidth for all smoother 对所有的拟合线用相同的带宽复选项，选中该复选项，在图形中所有的拟合曲线均用相同的带宽。

2. Spikes 散点连线选项卡见图 20-65。

(1) 在 Spikes 框中选择连线的方式。散点图的连线是从散点云团投射到某个点、某个坐标轴、某条线或某个面。使用连线可以帮助读者识别不同坐标轴上的数值，也可以帮助读者使用连线的长度比较各点的距离。

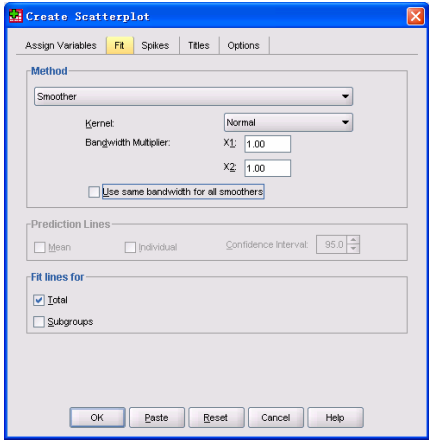


图 20-64 平滑拟合线选项卡

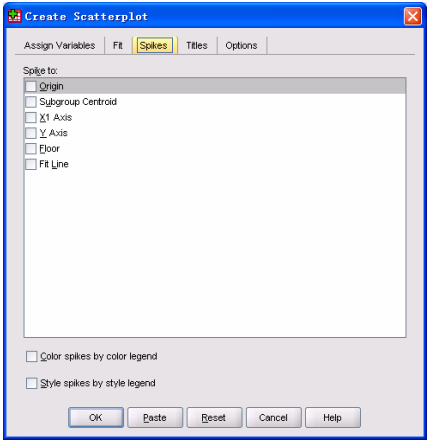


图 20-65 连线选项卡

- ① Origin 从散点云团投射到每个数据原始点的连线。
- ② Corner 从散点云团投射到某个焦点的连线。
- ③ Total Centroid 从散点云团投射到全部数据中心点的连线。
- ④ Subgroup Centroid 从细分组的散点云团投射到该分组数据中心点的连线。
- ⑤ X1 Axis 从散点云团投射到 X1 坐标轴的连线。
- ⑥ X2 Axis 从散点云团投射到 X2 坐标轴的连线。
- ⑦ Y Axis 从散点云团投射到 Y 坐标轴的连线。

⑧ Floor 从散点云团投射到  $X1$  和  $X2$  坐标轴平面的连线。

⑨ Fit Line 从散点云团投射到拟合线或拟合面的连线。

(2) 选中 Color spikes by color legend 项, 细分组散点连线的颜色与图例中分类变量的颜色匹配。

(3) 选中 Style spikes by style legend 项, 细分组散点连线的样式与图例中分类变量的样式匹配。

## 20.9 帕累托图

帕累托图 (Pareto Charts) 又可称为排列图或主次因素图。它作为改善质量管理活动中选择关键问题的一种工具, 由于关键的多数和次要的多数现象具有普遍性, 所以帕累托图也广泛应用于其他研究领域。

### 20.9.1 选择帕累托图类型

按 Anlyze→Quality Control→Pareto 顺序单击鼠标, 打开 Pareto Charts 条形图主对话框, 见图 20-66。

#### 1. 帕累托图图式

(1) Simple 简单帕累托图, 它对分类轴上的每一种类型的变量产生一条图, 并按各种因素发生次数的多少, 从左到右顺序排列, 帕累托曲线对分类轴上的每个变量值进行累加。

(2) Stacked 堆栈帕累托图, 是由分段条形图和帕累托图曲线构成的统计图。

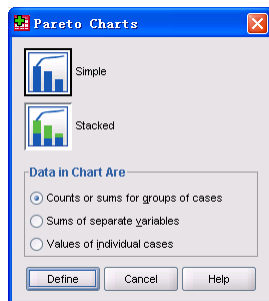


图 20-66 帕累托图主对话框

#### 2. 统计量描述模式

(1) Counts or sums for groups of cases 观测量分类数目或数值累加模式, 这种模式统计分类轴上的不同观测值数目, 或是对分类轴上观测值累加。

(2) Sums of separate variables 变量累加模式, 累加分类轴上每个变量。

(3) Values of individual cases 观测值模式, 对分类轴变量中的每一种观测值累加。

### 20.9.2 观测量分类数目或数值累加模式简单帕累托图

在帕累托图主对话框中选择 Simple 和 Counts or sums for groups of cases 项, 单击 Define 按钮, 展开定义观测量分类数目或数值累加模式简单帕累托图对话框, 见图 20-67。

1. Bars Represent 条图表达统计量栏, 条图可以表达两种不同类型的变量: 一种为字符型变量值, 另一种为数值型变量值。

(1) Counts 计数, 只适用于字符型变量。

(2) Sums of variable 变量累加, 适用于数值型变量, 被选定的变量在微框中显示。

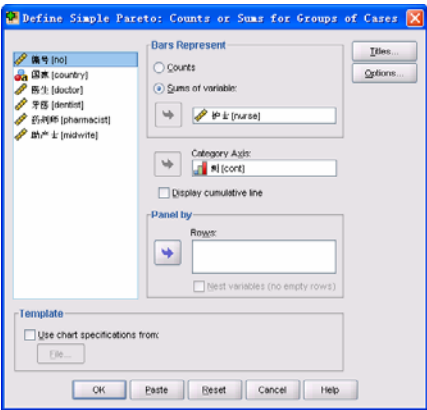


图 20-67 观测量分类数目或数值累加模式简单帕累托图对话框

2. Category Axis 选择分类轴变量框。

3. Display cumulative line, 系统默认为选定状态；选定此项，显示帕累托曲线（累积曲线）。

根据数据文件 data20-15，在 Bars Represent 框中选择 Counts，选择 cat 变量作为分类轴变量，生成切削刀质量帕累托图，见图 20-68(a)。通常把累计百分比分为三部分：0~80%表示主要因素（A 类），80%~90%表示次要因素（B 类），90%~100%表示一般因素（C 类）。

根据数据文件 data20-16，在 Bars Represent 框中选择 Sum of variable，选择 nurse 变量进入微框，并选择 cont 变量作为分类轴变量，生成的图形为各洲护士人数排列图，见图 20-68(b)。

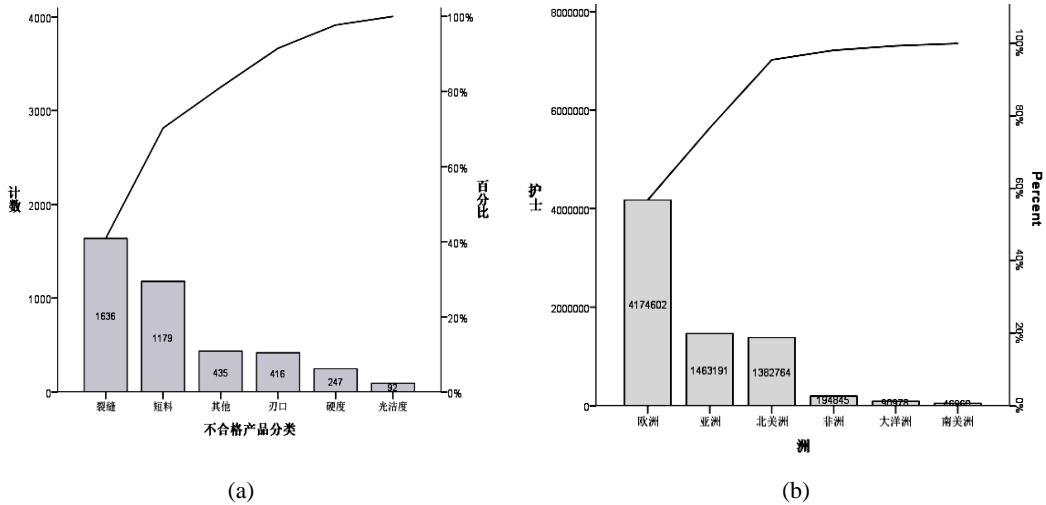


图 20-68 例图

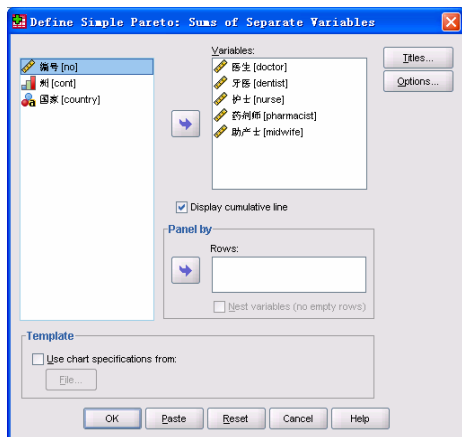
### 20.9.3 变量累加模式简单帕累托图

在帕累托图主对话框中选择 Simple 和 Sums of separate variables 项，单击 Define 按钮，展开定义变量累加模式简单帕累托图对话框，见图 20-69(a)。例题数据 data20-16，例图 20-69(b)为世界各地从事各种医疗保健人员的帕累托图。主要操作步骤为：

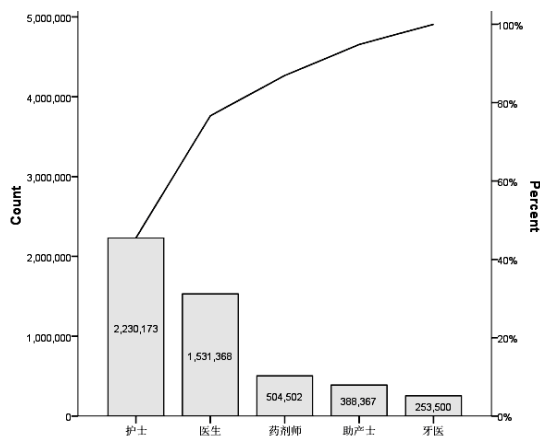
1. Variable 变量框：将 dentist、doctor、nurse、phar 和 widwise 变量选入此框。
2. Options 缺失值处理方式：选择 Exclude cases variable by variable 选项。
3. Display cumulative line 展示帕累托曲线（累积曲线）。

### 20.9.4 观测值模式简单帕累托图

在帕累托图主对话框中选择 **Simple** 和 **Variables of Individual Cases** 项，单击 **Define** 按钮，展开对话框，见图 20-70(a)。例题数据 data20-17。例图 20-77(b)为汽车空调蒸发器故障帕累托图，从图中可以了解到丢失螺钉是造成故障的最主要原因。

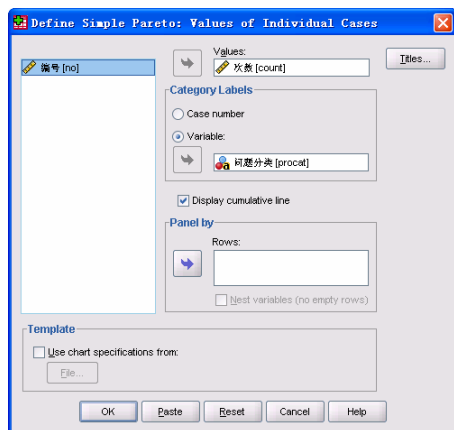


(a)

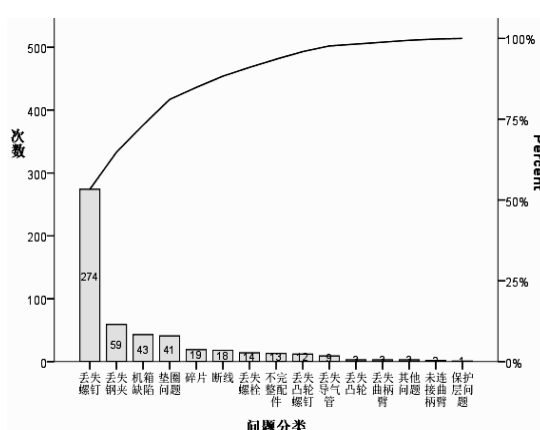


(b)

图 20-69 变量累加模式简单帕累托图对话框及例图



(a)



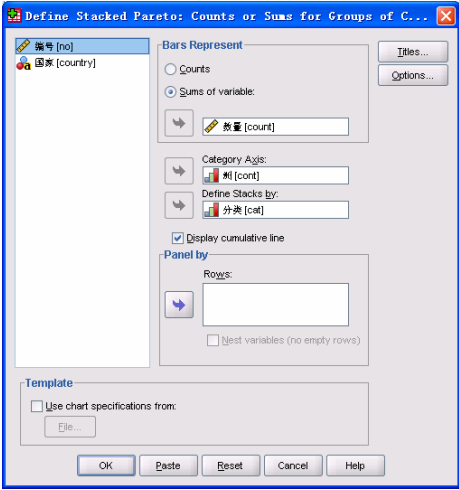
(b)

图 20-70 观测值模式简单帕累托图参数选择对话框及例图

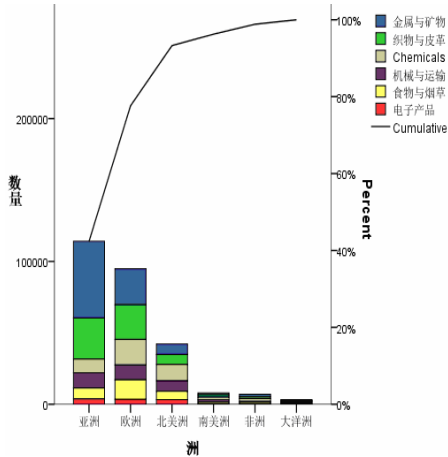
### 20.9.5 观测量数目或数值累加模式堆栈帕累托图

在帕累托图主对话框中选择 **Stacked** 和 **Counts or sums for groups of cases** 选项，单击 **Define** 按钮，展开对话框，见图 20-71(a)。例图 20-71(b)为各洲具有加工制造业工厂数量

帕累托图。数据 data20-18。主要操作是：在 Bars Represent 框中选择 Sums of variable 项；将 count 变量选入此框；选择 cont 作为分类轴变量；选择 cat 作为分段变量。



(a)

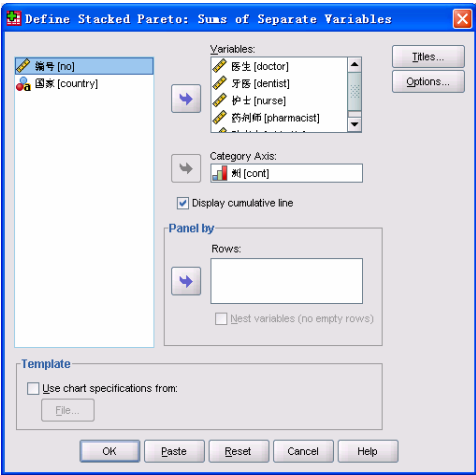


(b)

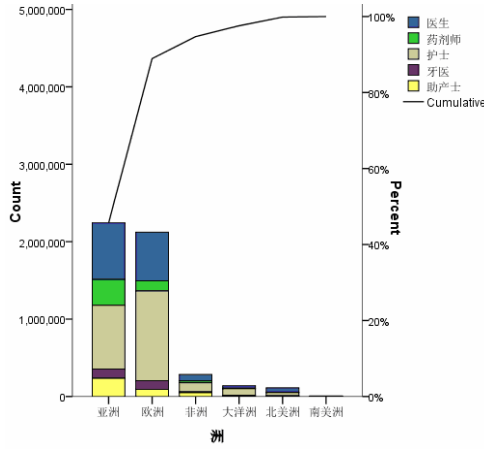
图 20-71 观测量分类数目或数值累加模式堆栈帕累托图对话框及例图

20.9.6 变量累加模式堆栈帕累托图

在帕累托图主对话框中选择 Stacked 和 Sums of Separate Variables 选项，单击 Define 按钮，展开对话框，见图 20-72(a)。本小节使用 data20-16 数据文件，例图 20-72(b)为各洲各类医疗人员总数帕累托图。



(a)

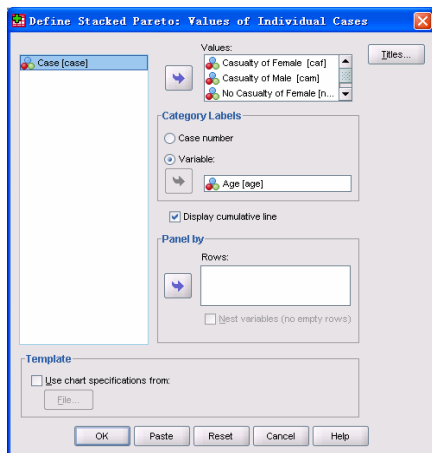


(b)

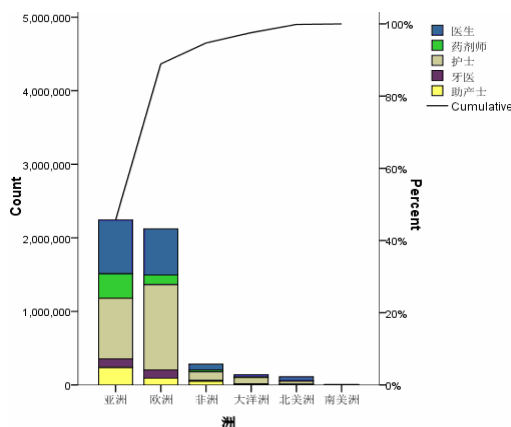
图 20-72 变量累加模式堆栈帕累托图对话框及例图

### 20.9.7 观测值模式堆栈帕累托图

在帕累托图主对话框中选择 **Stacked** 和 **Variables of Individual Cases** 选项, 单击 **Define** 按钮, 展开 **Define Stacked Pareto: Variables of Individual Cases** 对话框, 见图 20-73(a)。本节数据来自 data20-19 数据文件, 图 20-73(b)为各年龄段司机交通事故例数帕累托图。



(a)



(b)

图 20-73 观测值模式堆栈帕累托图对话框及例图

## 20.10 控制图

控制图 (Control Charts) 又称管理图, 它主要用于分析和判断生产工序是否处于稳定状态所使用的一种带有控制界限的统计图。虽然它始于产品质量的控制, 但以后推广到生产领域以外的许多方面, 诸如医学、金融等领域。控制图大致分为两类: 一类是计量值控制图, 另一类是计数值控制图, 在实际应用中, 这两类控制图常常是组合使用。

### 20.10.1 选择控制图类型

按 **Analyze**→**Quality Control**→**Control Chart** 顺序单击鼠标, 打开控制图主对话框, 见图 20-74。

#### 1. 控制图图式

(1) **X-bar, R, s** 项包括两种组合控制图, **X-Bar, R** 平均值-极差组合控制图和 **X-Bar, s** 平均值-标准差组合控制图。

(2) **Individuals, Moving Range** 单值-移动极差组合控制图。

(3) **p, np**, 包括 **p** 不合格品率和 **np** 不合格品数两种控制图。

(4) **c, u**, 包括 **c** 缺陷数控制图和 **u** 单位缺陷数控制图。

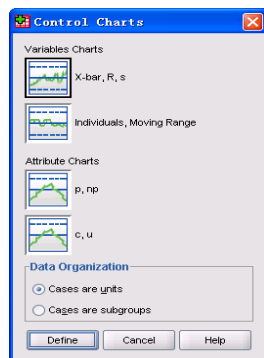


图 20-74 控制图主对话框



2. 数据编排方式（Data Organization）的选择

- (1) Cases are units 观测量组结构数据选择此项。如 data20-20 数据结构。
- (2) Cases are subgroups 变量组结构数据选择此项。如 data20-21 数据结构。

20. 10. 2 观测量组结构的平均值、极差、标准差控制图

在控制图主对话框中选择 X-bar, R, s 和 Cases Are Units 项，单击 Define 按钮，展开

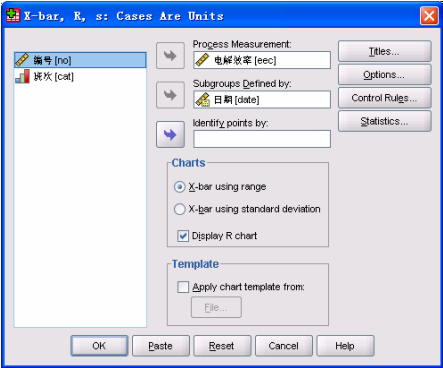


图 20-75 观测量组结构的 X-bar, R, s 对话框

X-bar, R, s: Cases are units 观测量组结构的平均值、极差、标准差控制图对话框，见图 20-75。根据 data20-20 中数据做图，见图 20-77。

1. Process Measurement 工序测量，本例中 eec 变量作为被测对象。

2. Subgroups Defined by 选定 date 变量为细分组变量，送入该框。不选择标识细分组变量，系统自动生成序号。

3. Charts 图形描述模式，有以下两种组合：

- (1) X-bar using range 平均值-极差控制图。
- (2) X-bar using standard deviation 平均值-标准差控制图。

这两个组合控制图的使用区别在于，前者用于细分组中样本数量较小的资料，后者用于细分组中样本数量较大（大于 10）的资料。本例选定 X-bar and range 项。

4. 单击 Options 按钮，出现 X-bar, R, s: Options 对话框，见图 20-76。

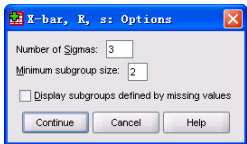
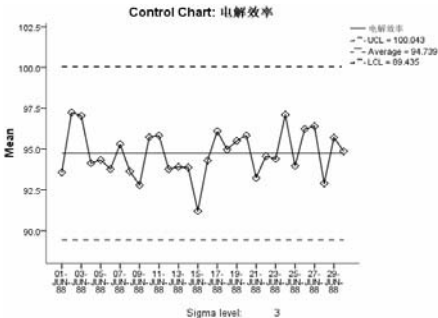
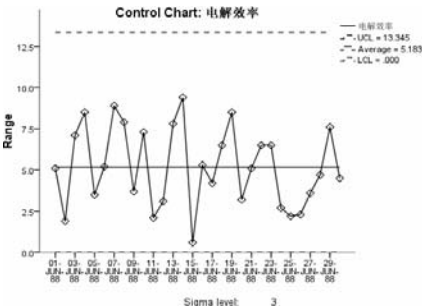


图 20-76 控制图选择对话框

- (1) Number of Sigmas 选择中心线上、下的标准差数值，默认值为 3。
- (2) Minimum subgroup sample size 指定细分组中最小样本数，默认值为 2。
- (3) Display subgroups defined by missing values 缺失值作为细分组显示。



(a)



(b)

图 20-77 例图

图 20-77(a)为每日三班电解工序的电解效率的平均值-极差控制图。图 20-77(b)是电解效率范围-极差控制图。

### 20.10.3 观测量组结构的单值-移动极差控制图

在控制图主对话框中选择 Individuals, Moving range 和 Cases are units 项, 单击 Define 按钮, 展开 Individuals, Moving Range 单值-移动极差控制图对话框, 见图 20-78。数据 data20-22, 图 20-79 为混凝土坍落度的单值-移动极差控制图。

1. Process Measurement 工序测量, 本例中 value 变量作为被测对象。

2. Subgroups Labeled by 项, 选定 no 变量为细分组的标识变量。

3. Chart 图形描述模式:

(1) Individuals and moving range 单值-移动极差控制图, 本例选定该选项。

(2) Individuals 单值控制图。

(3) Span 间距, 确定从哪个观测量开始作图, 默认值为 2。

4. Individuals and Moving Range: Options 单值-移动极差控制图设定。参见 20.10.2 小节。

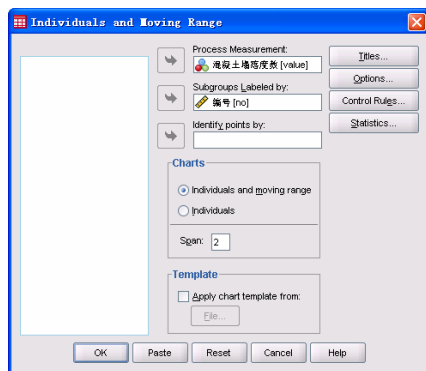


图 20-78 观测量组结构的单值-移动极差控制图对话框

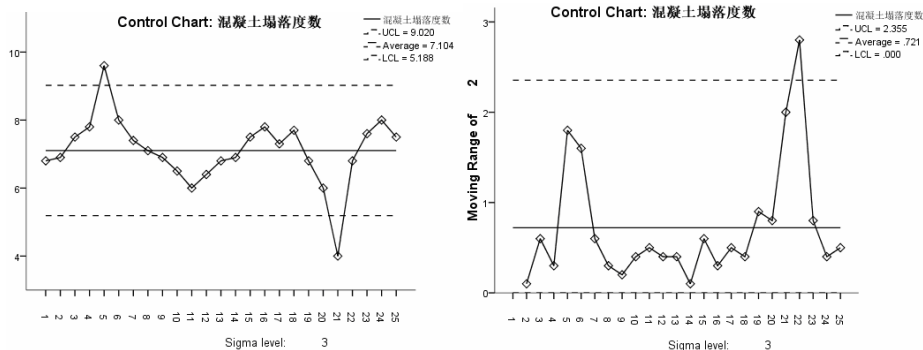
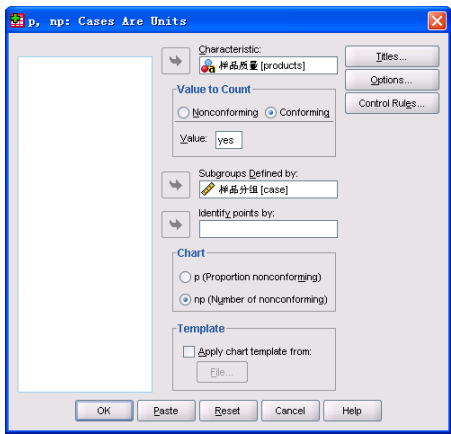


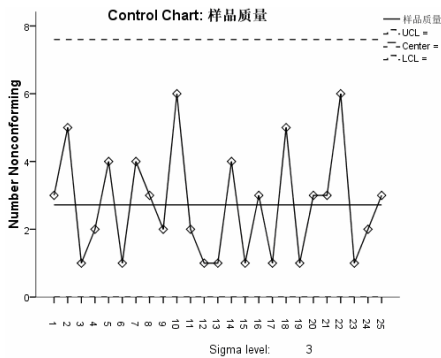
图 20-79 例图

### 20.10.4 观测量组结构数据的不合格品率、不合格品数控制图

在控制图主对话框中选择 p, np 和 Cases are units 项, 单击 Define 按钮, 展开 p, np: Cases Are Units 观测量组结构的不合格品率、不合格品数控制图对话框, 见图 20-80(a)。例题数据 data20-23, 图 20-80(b)为某种小螺钉不合格品数控制图。



(a)



(b)

图 20-80 观测量组结构的不合格品率、不合格品数控制图对话框及例图

1. Characteristic 质量测定，本例选择 products 变量作为被测对象。
2. Value to Count 变量值计数方式：
  - (1) Nonconforming 不合格品，本选项将计算不合格产品。
  - (2) Conforming 合格品，本选项将计算合格产品。
  - (3) Value 变量值属性，在 data20-20 数据文件中，将合格产品定为“yes”，不合格产品定为“no”，如果选择计算不合格产品数量，选用 Nonconforming 项，并在 Value 内录入 no。注意，这里输入的变量值应与被测变量中的变量值类型相同，例如被测变量为字符型，那么在这里的 Value 框内也要输入字符型变量。
3. Subgroups Defined by 细分组标识变量，本例选用 case 变量。
4. Chart 图形描述模式：
  - (1) p（Proportion nonconforming）不合格品率控制图。
  - (2) np（Number of nonconforming）不合格品数控制图，本例选定此选项。不论在 Value to Count 框选择了计算不合格品还是合格品选项，最后生成的图形都为不合格品数或不合格品率的控制图。
5. p, np: Options 不合格品率、不合格品数控制图设定，参见 20.10.2 小节。

20. 10. 5 观测量组结构的缺陷数、单位缺陷数控制图

在控制图主对话框中选择 c, u 和 Cases are units 项,单击 Define 按钮,展开 c, u: Cases Are Units 各观测量排列于同一变量的缺陷数、单位缺陷数控制图对话框,见图 20-81(a)。本小节使用 data20-24 数据文件,例图 20-81(b)为某医院每月出现危急外科手术的缺陷数控制图。主要操作为:选择 aes 变量作为被测对象送入 Characteristic 框中;选用 week 为细分组标识的变量送入 Subgroups Defined by 框。Chart 图形描述模式有两个选项:

(1) u (Nonconformities per unit) 单位缺陷数控制图。

(2) c (Number of nonconformities) 缺陷数控制图, 本例选定此选项。

对 c, u: Options 缺陷数、单位缺陷数控制图设定, 参见 20.10.2 小节。

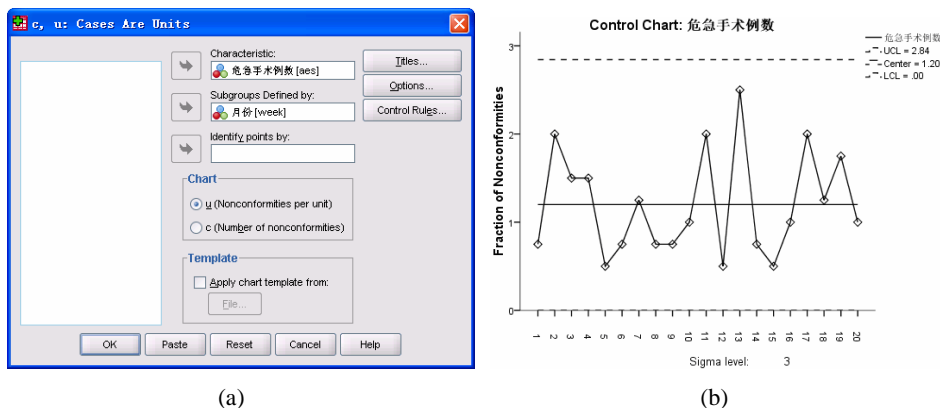


图 20-81 各观测量排列于同一变量的缺陷数、单位缺陷数控制图对话框及例图

## 20.10.6 变量组结构数据的平均值、极差、标准差控制图

在控制图主对话框中选择 X-bar, R, s 和 Cases Are subgroups 项, 单击 Define 按钮, 展开 X-bar, R, s: Cases Are Subgroups 变量组结构数据的平均值、极差、标准差控制图对话框, 见图 20-82。

使用 data20-25 数据文件, 图 20-83(a) 为 6mm±0.4mm 厚度钢板平均值控制图。图 20-83(b)为极差控制图。

1. Samples 样品测定, 至少选定两个或两个以上的数值型变量, 本例选择了 t1~t5 共 5 个变量。

2. Subgroups Labeled by 细分组标识, 本例选用 case 变量作为细分组标识变量。

3. Charts 图形描述模式, 参见 20.10.2 小节。

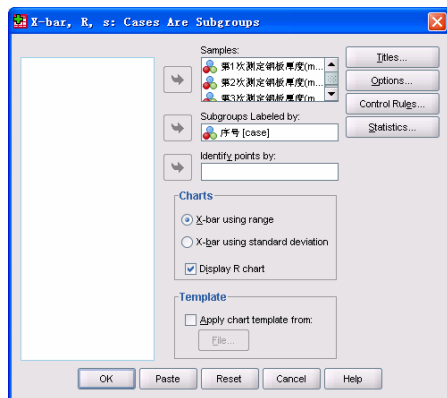


图 20-82 变量组结构数据的平均值、极差、标准差控制图对话框

## 20.10.7 变量组结构数据的不合格品率、不合格品数控制图

在控制图主对话框中选择 p, np 和 Cases Are Subgroups 项, 单击 Define 按钮, 展开 p, np: Cases Are Subgroups 变量组结构数据的不合格品率、不合格品数控制图对话框, 见图 20-84。本小节使用 data20-26 和 data20-27 数据文件, 生成例图见图 20-85。

1. Number of Nonconforming 不合格品数。

2. Subgroups Defined by 细分组标识。

3. Sample Size 样本量：

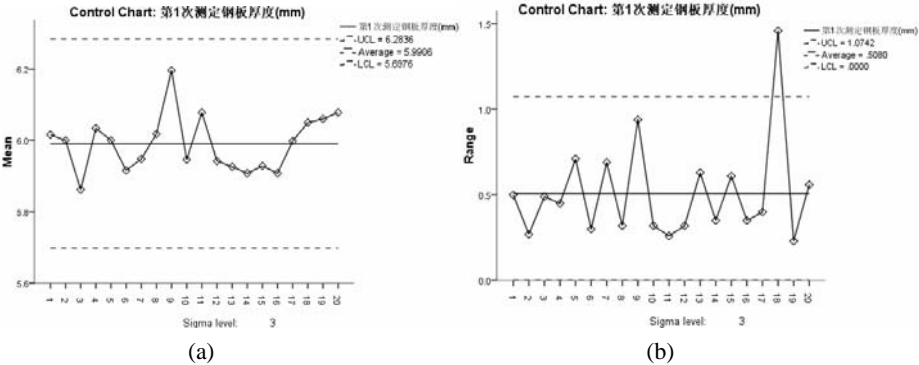


图 20-83 例图

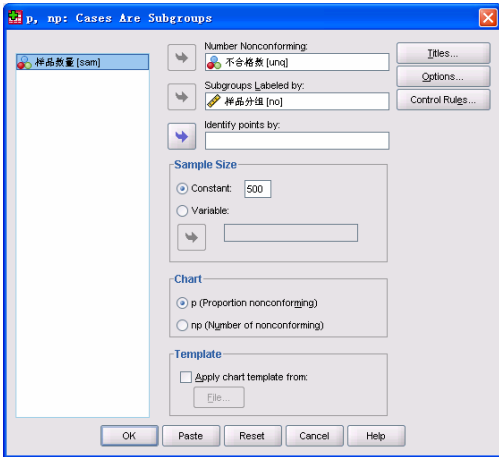


图 20-84 变量组结构数据的不合格品率、不合格品数控制图对话框

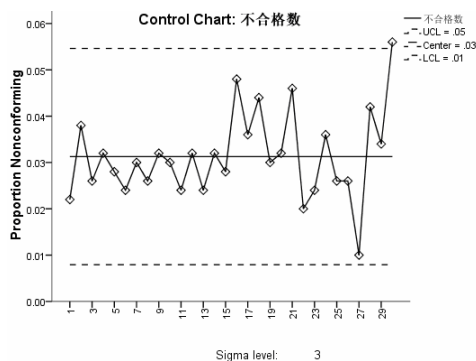
(1) Constant 细分组样本数量恒定，每个细分组样本量相同，选择此项并在微框中输入样本数。在 data20-23 数据文件中，由于每个细分组的样本数都是 500 个，故在此框内录入 500。

(2) Variable 变量确定样本数量，无论细分组样本数目是否相同，都可以通过录有每个细分组样本数量的变量来说明细分组样本数目。data20-24 数据文件中 sam 变量就是这样一个变量，所以用 sam 变量确定细分组样本数量。

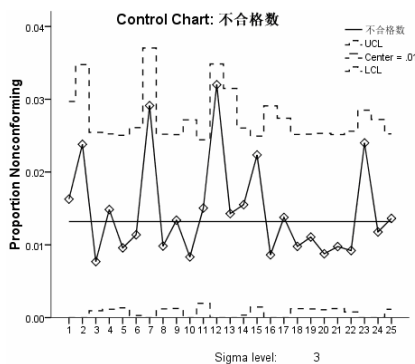
4. Chart 图形描述模式，参见 20.6.4 小节。

图 20-85(a)使用 data20-26 数据文件，在 Number of Nonconforming 框内选入 unq 变量；选用 no 变量作为细分组标识的变量；在 Sample Size 框中选择 Constant 项，并输入 500；最后在 Chart 框中选择 p 项，生成例图为某构件厂产品不合格品率控制图。

图 20-85(b)是使用 data20-27 数据文件, 在 Number of Nonconforming 框内选入 unq 变量; 选用 no 变量作为细分组标识变量; 在 Sample Size 框中选择 Variable 项, 在微框内输入 sam 变量; 最后在 Chart 框中选择 p 项, 生成例图为某种小螺钉不合格品率控制图。



(a)



(b)

图 20-85 例图

## 习 题 20

1. 绘制统计图形有哪些基本要求?
2. 以 data20-28 为原始数据, 绘制以下图形:
  - 男性和女性期望寿命对比条图
  - 不同气候地区的国家数量图
  - 不同地区平均国民总产值交互图形
  - 世界上不同宗教所占百分比圆图

图 21-2 图形组成说明

## 21.2 编辑平面统计图

### 21.2.1 图形编辑途径

在输出观察窗口中产生图形后,为了进一步探查数据或增强视觉效果,需要在 Chart Editor 图形编辑窗口编辑所生成的图形。

#### 1. 编辑图形的三种途径

要编辑生成的图形基本操作是双击它,进入图形编辑状态,即在图形编辑窗口显示待编辑的图形,如图 21-3 所示。同时打开道具窗如图 21-4 所示。

图 21-5 图形编辑窗标题栏下面,自上至下分别为功能菜单、编辑工具、选择工具、元素工具和格式工具。窗口右下角还有状态栏。是否在窗口中显示这些工具,可以在 View 菜单项中选择。在编辑器窗口使用菜单项和工具栏中的工具对图形进行编辑是对图形编辑的第一种途径。这里许多菜单中的命令已经在前面的章节介绍过。

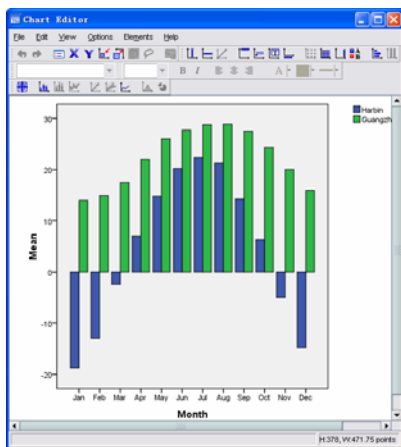


图 21-3 图形编辑窗



图 21-4 图形编辑道具窗



图 21-5 主菜单和 4 种工具

在打开编辑器同时如果没有打开道具窗口,可以按 Edit→Properties 顺序单击菜单项打开道具窗。也可以在外框内空白处单击右键,选择第一项 Properties Window 打开道具窗。在编辑窗口的图形上选择了一个要修改的图形元素,道具窗的内容发生相应的改变。其选项卡和各选项卡中的编辑方法与选中的图形元素对应。很多操作可以在道具窗中完成。使用道具窗各种选项卡中的功能是编辑图形的第 2 种途径。

鼠标右键单击待编辑的图形元素时展开右键菜单,其中包括各种编辑该图形元素和有关的元素组的功能。图形不同、选择的图形元素不同,右键菜单的内容也各不相同。选择其中的功能项,也可以对图形元素进行具体的编辑。这是编辑图形的第 3 种途径。

这三种途径是相通的。但都必须是在打开编辑器在编辑窗中才能实现编辑功能。

本节主要介绍与图形编辑有关的命令和工具和道具窗和右键菜单的操作。




## 2. 选择编辑对象

(1) 图形元素的选择。要对图形元素进行编辑,必须首先选择它。要编辑的图形元素有时是单个元素,例如选择坐标轴,有时是一组,例如轴上刻度的标签。被选择的图形元素被彩色框框住。选择方法很多,介绍几种常用方法:

① 单击选择。例如选择坐标轴,选择饼图的所有扇面、选择天下的数据区。

② 单击两次选择,往往用于选择并列元素组中的一组元素,例如选择双变量条形图中的一个变量的一组条。

③ 右键菜单选择,用于指定元素组中的一个成员时。例如要选择条形图中的一个条。鼠标右键单击其中一条,在右键菜单中选择 **Select→ This Bar**。

④ 用套索选择几个图形元素,如在散点图中选择离群点,可以单击套索工具,用套索光标对要选择的对象画封闭曲线。

(2) 文字的选择与移动、放缩。选择文字方法与图形元素的选择方法相同。选择后,四周出现带有八个方块的框,鼠标指针置于方块上按住鼠标键可以放缩套住的文字;鼠标光标置于框的边缘,鼠标光标变成四个箭头,可以按住鼠标键,移动图形元素到新位置。见示意图 21-6。

在如图 21-4 所示的道具窗 **Chart Size** 选项卡中可以精确地放缩图形大小,单击 **Height**、**Width** 旁的上下箭头,改变图形外框的高度和宽度。要想保持高、宽比,按比例放缩,选择

**Maintain aspect ratio**。放缩外框以内的各图形元素,但文字部分,例如主副标题、轴标题、轴刻度标签、图例不变化。选择 **Resize elements when new elements are added/removed** 当新元素加入和移出时重新自动调整图形大小。



图 21-6 使用鼠标移动、放缩

### 21.2.2 改变图形构成

为说明编辑方法,先做出条形图。数据 data21-01 是 12 个城市 1985~1994 年各月的气温数据。单击 **Graphs→Legacy Dialogs→Bar**,在 **Bar Charts** 窗口选择 **Cluster/Summaries of separate variables→Define** 按钮打开 **Define Clustered Bar: Summaries of Separate Variables** 对话框。在源变量框中选择 **Month** 月份作为 X 轴变量送入 **Category Axis** 框中;把广州和哈尔滨的气温 **guangzhou**、**harbin** 两个变量送入 **Bars Represent** 框中。

单击 **Titles** 按钮,输入图形标题,Line1: “月均气温比较”;副标题 Subtitle: “1985~1994”单击 **Continue**,主对话框中单击 **OK** 按钮,生成图形如图 21-7(a)所示(去掉了标题)。

#### 1. 图形转换

图形转换必须要有充足的数据。系统自动识别可以转换的图形,在 **Properties** 道具窗口 **Variables** 选项卡的 **Element Type** 下拉列表中加黑的图形就是当前图形可以转换成的目标图形。转换成线图和散点图的结果见图 21-7(b)、(c)。操作方法是:

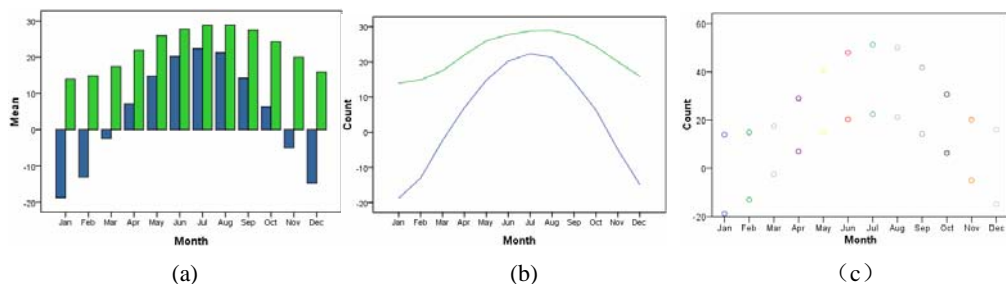


图 21-7 图形转换

双击该条形图，在道具窗 Variables 选项卡的 Element Type 下拉列表中分别选择 Interpolation Line、marker，每选择一个，单击 Apply 一次。转换成的线图和散点图如图 21-7(b)、(c)所示。

注意，变化后的图形类型不一定能很好表达数据，不一定能方便观察。例如，本例若要转换成饼图，就不易直接观察。所以要注意选择最后的转换结果。

## 2. 图形转置

对有 X、Y 轴的平面图形可以进行转置，即把直角坐标系旋转 90 度。它与改变源变量与自变量的角色还有区别。例如对曲线图形来说，后者需要重新进行拟合。

图形转置的方法很简单。在右键菜单中选择 Transpose chart 即可得到转置后的图形，见图 21-8。

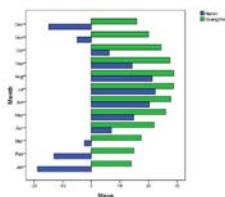




图 21-8 转置结果

## 3. 在图形中增加值标签

读者可以显示条形图中的条、圆图中的扇、线图上的点、箱线图的中线所代表的数值、百分比，或散点图和箱线图各个观测量的数值。

首先选择要显示的数值的图例，在图 21-9(a)图例中选择“Harbin”，鼠标右键菜单中选择 Show Data Label，所有的哈尔滨的平均气温数值标出。单击右键菜单中的 Hide Data Label，值标签消失。

选择值标签，道具窗口增加 Data Value Label 选项卡，在这个选项卡中 Displayed 框中是已经显示的标签， Not Displayed 框中是还没有显示的，但可以加上的标签内容。例如我们选择 Month，送入上框，单击 Apply 按钮，哈尔滨的气温条上增加了月份值，见图 21-9(b)。增加值标签的道具窗如图 21-10 所示。

单击工具栏中的“”工具，光标变成“”形状时，表示激活数据识别模式，用光标单击某个图形中的条、点、线等即可显示/隐藏数值标签。这个工具的方便之处在于可以逐个增加标签，也可以逐个隐藏已经加上的标签。

## 4. 增加图形组成

图形生成后可以对图形继续修饰。例如，增加对图形的解释，对一些变量或数值注释，画出参考线、拟合线等。图 21-11 是右键菜单，增加图形元素的选择项有：Add X Axis

Reference Line 增加 X 轴参考线、Add Y Axis Reference Line 增加 Y 轴参考线、Add Title 增加标题、Add Annotation 增加注释（内框内、数据区中）、Add Text Box 增加文本框（外框内、内框外）、Add Footnote 增加注脚。

图 21-12 是增加图形元素：X 轴参考线、Y 轴参考线、注释、文本框、注脚的结果。

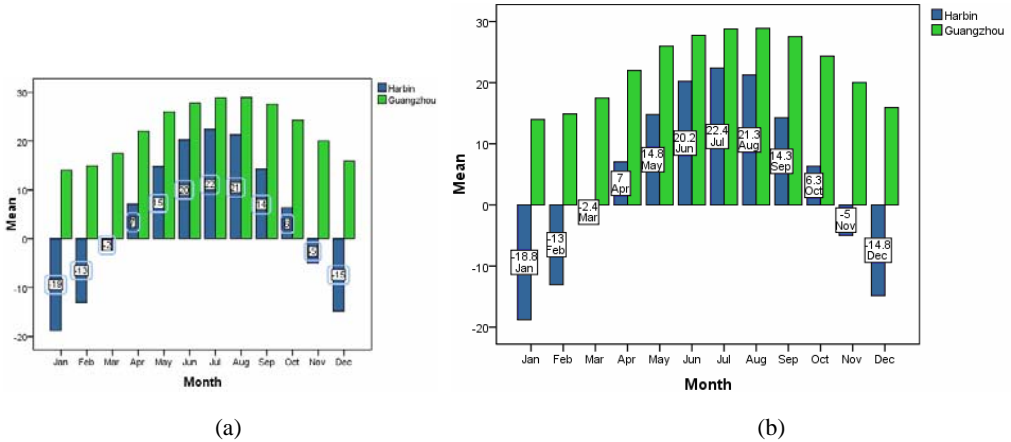


图 21-9 数值标签

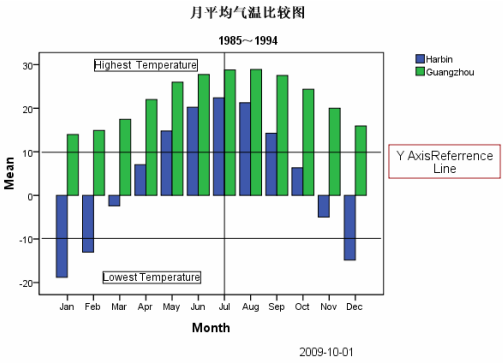
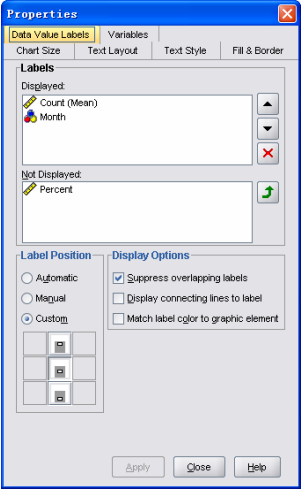


图 21-10 数值标签选项卡

图 21-11 右键菜单

图 21-12 增加新图形元素的条形图

5. 显示派生轴、图例、线图标记点见图 21-13。

右键菜单中选择 Show Derived Axis，显示派生轴，再选择 Hide Derived Axis 顺序单击鼠标，隐藏派生轴。

在原图中有图例。在右键菜单中选择 Hide Legend 隐藏图例。选择 Show Legend 显示图例。

编辑线图时，右键菜单中选择 **Show Line Markers** 显示线图标记点。相反，选择 **Hide Line Markers** 隐藏线图标记点。

6. 改变分类在道具窗 **Categories** 选项卡中进行。例如对条形图、散点图等改变图中分类变量各类的顺序、增加或减少某一类的条或点，对圆图减少增加某一类的扇面等。

(1) **Variables** 下拉列表中可以指定变量，也可以指定图例 **Legend**。Variable 下拉列表指定一个设置一个。图 21-14(a)为对 **Legend** 图例所列国家分类的操作，减去一个图例。图 21-14(b)是对分类变量 **mon** 月份的操作，去掉单月的图条。

(2) **Collapse (sum) categories less than** 在后面输入百分比值，将小于设置数值的元素合并为一类。如输入 10，凡是图中数值总和小于 10% 的分类，合并为默认名为“other”的新分类。

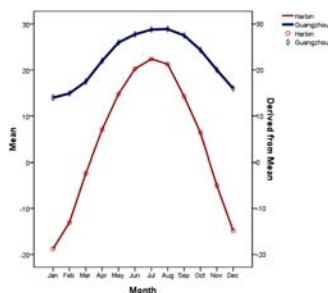
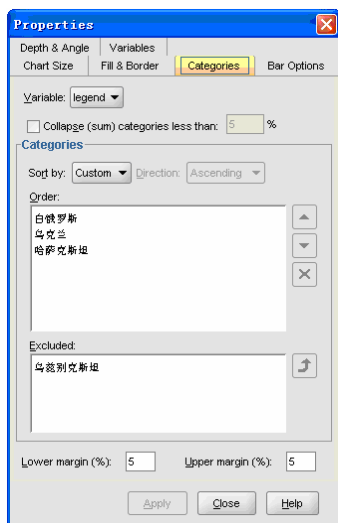
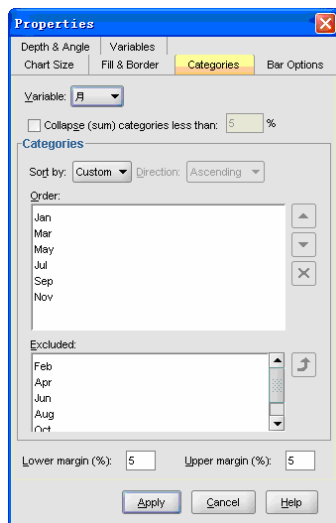


图 21-13 派生轴、图例和标记点



(a)



(b)

图 21-14 Categories 选项卡对 2 个分类变量的设置

(3) **Categories** 分类栏显示分类变量的各类排序的方法和方向：

① **Sort by** 下拉列表中选择分类的排序是按：**Value** 分类变量值、**Label** 值标签、**Statistics** 统计量、**Custom** 读者定义的顺序排序；排序结果显示在 **Order** 栏中。

② 选择前 3 种排序分类值的方法，需要在 **Direction** 下拉列表中选择按 **Ascending** 升序、**Descending** 降序排序。

③ **Order** 显示分类轴上显示的分类值, 如果选择了自定义, 通过上下箭头按钮调整分类值的顺序。选中某分类值, 单击叉子按钮将其移到 **Excluded** 框中, 在分类轴上不再显示该分类值的图形。

④ **Excluded** 框中是被剔除的分类值。选中一个, 单击向上按钮, 该值送回 **Order** 框。

(4) 分类轴两侧留白选择项 **Lower margin (%)**、**Upper margin (%)** 表示在分类轴左侧、右侧留出的空间占整个分类轴的百分比。

图 21-15(a) 是独联体 4 国家失业月数据条形图, 上图为原图, 下图去掉一个国家和去掉单月数据条的结果。

图 21-15(b)的上图是原图, 下图是将小于 8% 的扇面合并, 按统计量排序 10 月、11 月的扇面互换了位置的结果。

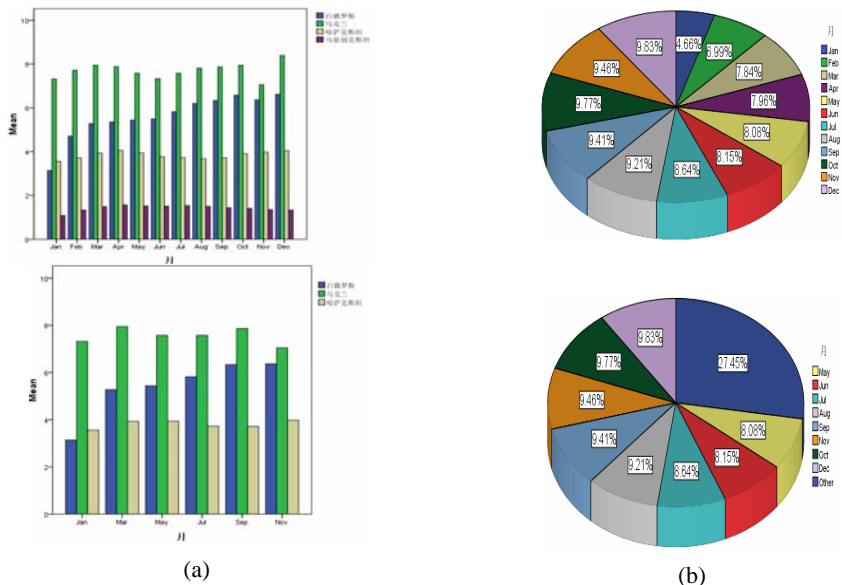


图 21-15 原图与效果比较

### 21.2.3 图形与文字修饰





对图形的修饰在道具窗中的 **Fill & Border** 选项卡中进行, 见图 21-16(a); 对文字的修饰在道具窗中的 **Text Style** 和 **Text Layout** 选项卡中进行, 分别见图 21-16(b)、图 21-16(c)。如果选择了要修饰的图形元素或文字, 没有出现道具窗, 按 **Edit→Properties** 打开道具窗。

#### 1. 填充与边框

填充功能是对图形中的整体或选中的区域进行填色或增加底纹。边框是对选定的区域增加线框, 改变边框的线型、粗细、颜色。被选定区域可以包括全部图形、图形内框区、图例框、文本框、注释框等区域, 还包括条图、面积图、极差图、圆图、箱线图、

误差条图、直方图等。

(1) **Preview** 预览当前的和单击 **Apply** 按钮后实现的填充颜色、底纹、框线的样式。

(2) **Color** 颜色栏: **Fill** 框中显示填充颜色, **Border** 框中显示边框颜色, **Pattern** 下拉菜单中选择填充底纹。在调色板中, 选择  填充黑色, 选择  填充白色, 选择  填充透明色。内框内背景色在创建图形后显示为灰色, 单击  填充效果较好。

(3) **Border Style** 边框样式栏: **Weight** 框中选择线条粗细, **Style** 框中选择线型, **End Caps** 框中选择虚线类线型的每一段线两端的形状。

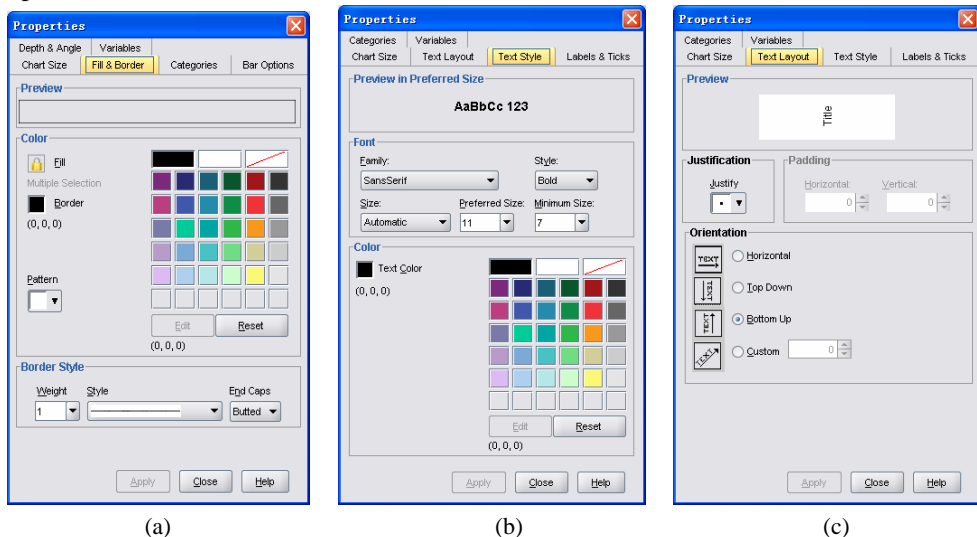


图 21-16 修饰图形元素、文字的道具对话框

## 2. 修饰文字

图形中的文字包括 **Text Box** 键入的文本、**Title** 图形标题、**Subtitle** 子标题、**Footnote** 脚注、**Axis Title** 轴标题、坐标 **Axis Value Label** 轴数值标签、**Legend Title and Item** 图例标题等。

### (1) Text Style 选项卡

① **Preview Preferred Size** 以首选大小显示当前文字式样, 选择了字体、字号、颜色后的预览和单击 **Apply** 后实际的文字式样。

② **Font** 栏, 在 **Family** 下拉菜单中选择字体, 在 **Style** 下拉菜单中选择是否加粗、倾斜或同时加粗倾斜。在 **Size** 下拉列表中选择字号; **Preferred Size** 下拉列表中选择首选字号; **Minimum Size** 下拉列表中选择最小字号。如果图形不是太小, **Size** 中选择 **Automatic**, 显示的字号与整个图成比例, 首先尝试使用首选字号, 再小也不会小于最小字号。

③ **Color** 颜色栏中选择字的颜色, 显示在 **Text Color** 中。

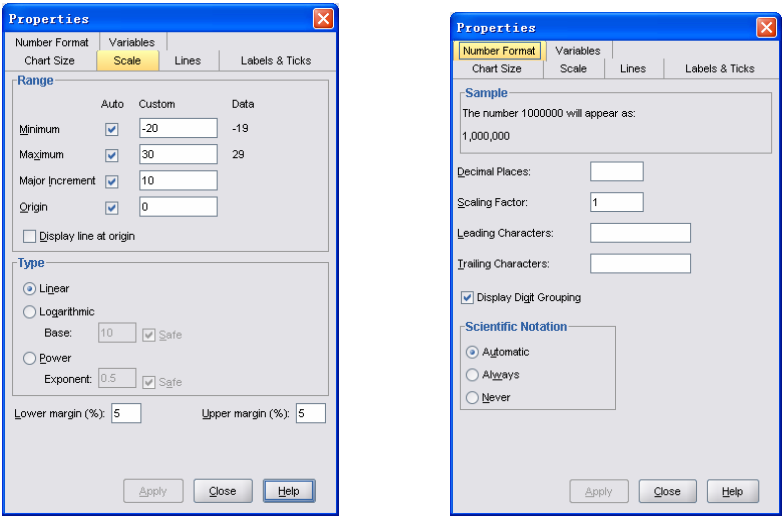
(2) **Text Layout** 选项卡中设置文字布局, 可以在预览栏中看到选择结果。

① **Justification** 栏 **Justify** 下拉列表中选择文字对齐方式；**Padding** 栏设置文字在它的框架范围中与框架的距离，包括水平和纵向距离。如果对齐方式不是中心对齐，距离设置不当有可能显示不出文字。

② **Orientation** 栏选择文字排列方向。分水平、纵向的自上而下和自下而上、自定义倾斜角度的 4 种方式。

21.2.4 坐标轴的编辑

在图形编辑窗口中，选中坐标轴，打开 **Properties** 道具窗。如果没有显示道具窗，按 **Edit→Properties** 顺序单击鼠标，打开道具窗口。坐标轴编辑可能使用到的选项卡有如图 21-17(a)所示的刻度选项卡 **Scale**、如图 21-17(b)所示的数值格式选项卡 **Number Format**、如图 18(a)所示的线型选项卡 **Line** 和如图 21-18(b)所示的 **Labels & Ticks** 标签与标记选项卡。



(a) (b)  
图 21-17 坐标轴编辑的刻度和数值格式选项卡

1. **Scale** 选项卡，如果选择的坐标轴变量是尺度变量，从道具窗中选择此选项卡。

(1) **Range** 栏中设置坐标轴刻度范围、跨度和刻度起始位置。

① 选择刻度范围和跨度。它们之间的关系是：

	Auto 自动设置	Custom 自定义	Data 数据
Minimum	系统指定最小值	用户指定最小值	实际数据最小值
Maximum	系统指定最大值	用户指定最大值	实际数据最大值
Major Increment	系统指定主刻度跨度	用户指定主刻度跨度	
Origin	系统指定起始值	用户指定起始值	



② Display line at origin 在 origin 显示直线。例如，图 21-12 在 Y 轴 0 处显示横线。

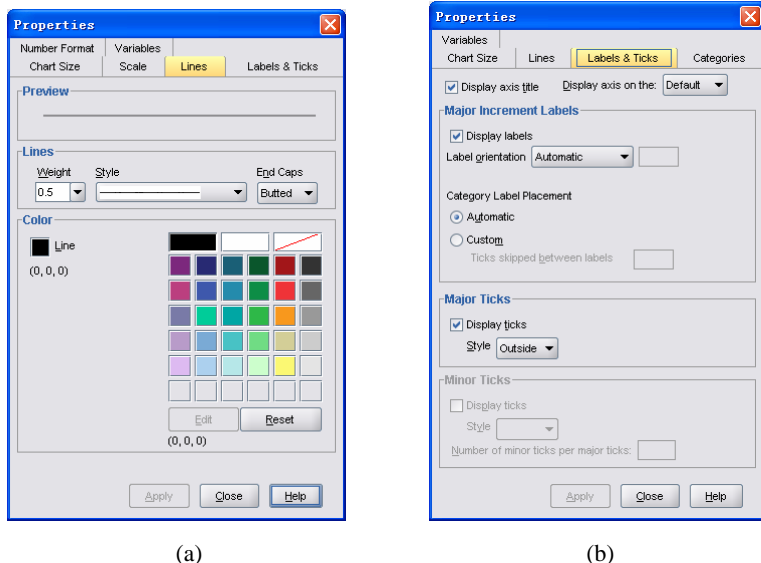


图 21-18 坐标轴编辑的选项卡

(2) Type 栏选择坐标轴变换的方法

① Linear 显示线性的未转换的刻度，为便于得出统计结论对刻度进行转换。

② Logarithmic 显示对数转换的刻度，在 Base 后输入对数的底，其值必须大于 1。如果选择了 Safe，则不是以  $\log(y)$  做刻度，而是对原刻度的绝对值取对数，加上符号。

③ Power 显示指数幂刻度，在 Exponent 后输入指数，默认值 0.5，即开平方。选择了 Safe 则显示安全的刻度，即在原刻度处的数值取指数，而不是对轴变量取指数再作图。

④ Lower margin、Upper margin 设置图形数据区元素的上下留白的百分比，默认 5%，可以输入的值范围 0~50。如果输入 50 则看不到图形。

2. Number Format 数值格式选项卡，见图 21-17(b)。如果选择的坐标轴变量是尺度变量，从道具窗中选择此选项卡。如果没有出现道具窗，可以从右键菜单中选择。

① Decimal Places 参数框中输入刻度标识的小数点位数。

② Scaling Factor 参数框指定比例因子，刻度轴上的每个值除以换算系数，例如刻度值为 1,000,000、2,000,000... 可以用 1、2... 来代替，同时“millions”加到轴的标题上。

③ Leading Characters 指定刻度标识的第一个字符，例如“\$”。

④ Trailing Characters 指定刻度标识的最后一个字符，例如“%”。

⑤ Display Digit Grouping 指定使用千位分节号。

⑥ Scientific Notation 科学计数栏：Automatic 自动确定是否使用、Always 一直使用、Never 从不使用科学计数方式。



3. 坐标轴线型选项卡见图 21-18(a)可在预览栏中看到选择的实际效果。

(1) Lines 栏中, 在 Weight 下拉列表中选择线的粗细; 在 Style 下拉列表中选择线型; 在 End Caps 下拉列表中选择非实线线型的每一段两端的形状。

(2) Color 栏中确定坐标轴的颜色。

4. Labels & ticks 选项卡见图 21-18(b)。设置轴上的值标签和刻度线的属性。

(1) Display axis title 显示坐标轴标题, 默认显示。

(2) Display axis on the\_ 设置坐标轴显示的位置。默认 X 轴在下边, Y 轴在左边。选择 Opposite 坐标轴显示到默认位置的对面。

(3) Major Increment Labels 主刻度标签栏

① Display labels 显示坐标轴刻度标签, 选中该项, 激活 Label orientation 刻度标签编排框。可选的排列方式: Automatic 自动; Horizontal 水平; Vertical 垂直; Diagonal 对角; Staggered 交错; Custom Degrees 自定义角度, 在框中输入旋转的角度。

② Category Label Placement 分类轴刻度标签放置, 以下两个选项只对分类轴有效。

- Automatic 系统自动放置。

- Custom 自定义放置。Ticks skipped between labels, 指定跳过某些刻度, 在数值框输入的数字, 决定跳过的标签数。例如, 在分类轴上有 A~L 共 12 个分类刻度标签, 在数值框中输入“2”, 最后在分类轴上只显示 A、D、G、J 共 4 个标注。

(4) Major ticks 主刻度标记

① Display ticks 显示主刻度。选中该项, 在 Style 下拉列表中选择显示方式 Outside: 刻度标记在坐标轴外, Inside: 在坐标轴内, Through: 刻度标记穿越坐标轴线。

② Minor Ticks 次刻度标记栏: 选择 Display ticks 显示次刻度标记。在 Style 下拉列表中选择显示方式。选择项含义与主刻度标记相同。

③ Number of minor ticks per major ticks 后设置每两个主刻度之间的次刻度数。

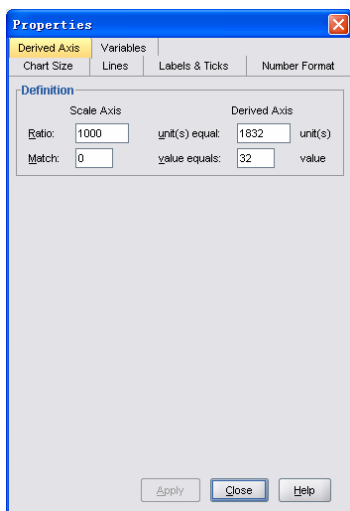
5. 派生 Y 轴的修饰 (派生方法见 21.2.2 小节)

选中派生出的 Y 轴, 在道具窗中选择 Derived Axis 派生轴选项卡, 见图 21-19(a)。在 Definition 栏内定义派生轴与尺度轴的对比关系。

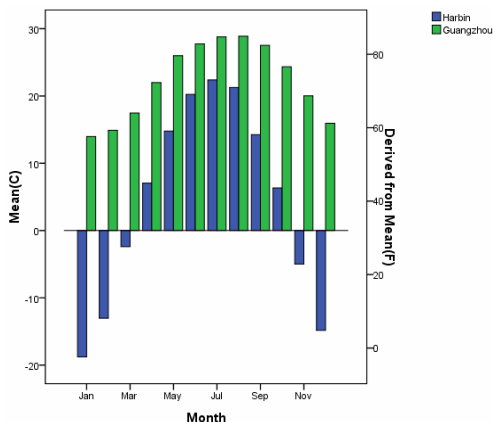
(1) Ration: \_\_unit(s) equal \_\_unit(s) 行上定义原 Y 轴单位与派生 Y 轴单位的比率, 例如 Ration: 1000 unit(s) 1832 unit(s), 意为原 Y 轴 1000 个单位相当于派生 Y 轴 1832 个单位, 两个数字框中必须输入正整数。此外, 在指定了比率之后, 还应该考虑增量的大小。

(2) Match: \_\_value equal \_\_value 原 Y 轴上某数值与派生 Y 轴上某数值相对应, 例如 Match: 0 value 32 value, 即尺度轴 0 与派生轴 32 相对应, 这种对应关系不一定在图中显现。以上设置显示在图 21-19 上, 结果是派生轴上显示华氏温度的近似值。

图 21-19(b)为哈尔滨—广州的月平均气温条形图, 原 Y 轴是摄氏度, 派生轴为华氏度刻度。X 轴是分类轴刻度标签按自定义排序; 主刻度向外, 隔 1 个值显示一个, 显示轴标题, 横轴两端各留 10% 的空间; Y 轴的原点 (摄氏 0 度) 显示一直线。



(a)



(b)

图 21-19 派生轴选项卡及编辑结果

### 21.2.5 图条的修饰

当生成的图形为条、箱线、误差条、垂线、极差和高低图时，可以对图条进行修饰。图条的修饰在两个选项卡中进行：**Bar Option** 选项卡和 **Depth & Angle** 选项卡。分别见图 21-20(a)、(b)。以条形图为例说明对类似的条线修饰的方法。

选中图形中某个图例，打开道具窗，选择 **Bar Options** 修饰选项卡，见图 21-20(a)。

#### 1. Bar Option 选项卡

(1) **Width** 宽度栏：移动游标或在参数框中输入图条的宽度占系统给出范围的百分比。**Bars** 调整条图组内间距百分比，**Cluster** 调整条图组间间距百分比。

(2) **Link the box, median line, and error bar widths** 在箱图中选择箱体、中线或在误差条图中的误差条，移动游标则调整这些元素的宽度。

(3) **Scale boxplot and error bar with based on count** 在箱线图和误差条图中，选中此项，根据分类变量各分类中所含观测量的多少决定每个图条的宽窄。

(4) **Boxplot and Error Bar Style** 栏，选择箱线图和误差条图的外伸线的样式。

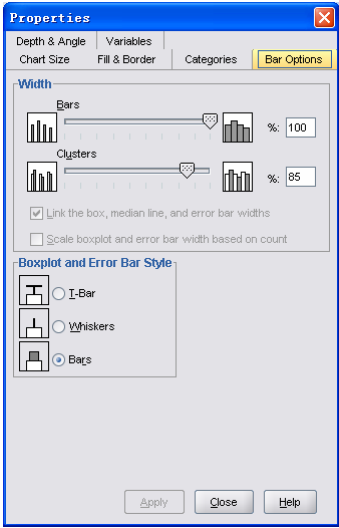
#### 2. 平面效果和立体效果转换

选中某个图例，在道具窗选择 **Depth & Angle** 深度和角度选项卡，见图 21-20 (b)。

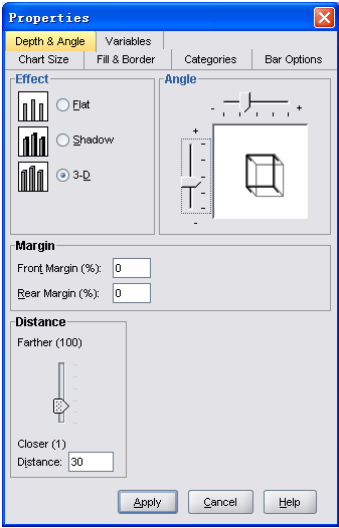
(1) **Effect** 栏：选择图形效果。**Flat**：平面图，**Shadow**：阴影图，**3-D**：立体图。

(2) **Angle** 栏：拖动标尺，选择阴影图和立体图的水平和垂直角度。通过调整角度，立体图表现出不同的深度。

(3) **Margin** 栏：**Front Margin(%)**和 **Rear Margin(%)**框，分别选择立体图前后两侧留白空间占内框的百分比。



(a)

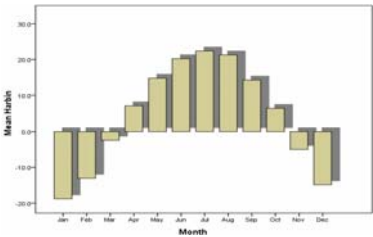


(b)

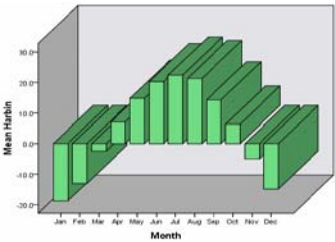
图 21-20 图条修饰选项卡、深度和角度选项卡

(4) Distance 栏：改变立体图形视觉上的远近，直观形成图形大小。在 Distance 框中输入距离数值，确定图形的大小。

图 21-21(a)是阴影条形图，(b)是 3-D 条形图。



(a)



(b)


图 21-21 阴影与 3-D 条形图

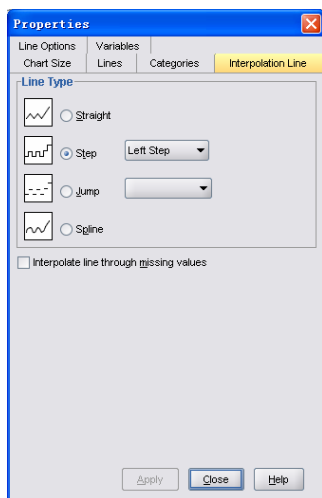
### 21.2.6 图线的编辑

修饰图线使用道具窗的 Interpolation Line 连线选项卡，见图 21-22(a)；Line Options 线形选项卡，见图 21-22(b)。选定图线，打开 Properties 道具窗。

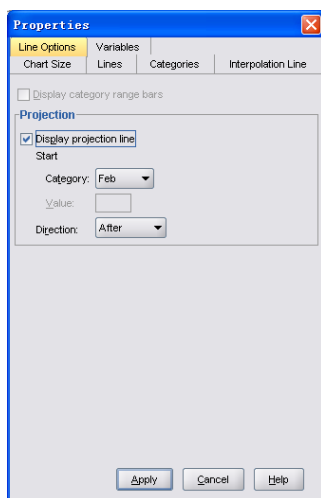
1. Interpolation Line 连线选项卡见图 21-22(a)，确定连线方式。系统默认在线图上连接各点的方式为折线，连线还可以应用在散点图、高低图、误差条图等图形上。

(1) Line Type 栏中选择连线方式。

 Straight 折线：各数值点之间用直线相连，此为默认方式。



(a)



(b)

图 21-22 连线选项卡和突出线选项卡

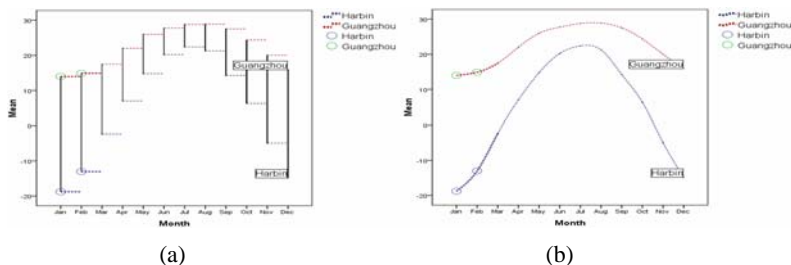
**Step 阶梯线：**在图中水平线穿过每个数值点，垂线连接左右数值点的水平线。数值点在水平线上有三种位置可选：Left step 左侧，Center step 中点，Right step 右侧。

**Jump 跳跃线：**在图中只用水平线穿过每个数值点，各点间没有连接线。数值点在水平线上有三种位置可选：Left jump 左侧，Center jump 中点，Right jump 右侧。

**Spline 曲线：**一个三次曲线从左至右通过数值点。

(2) Interpolation line through missing values，连线通过缺失值。

2. Line Options 选项卡见图 21-22(b)，例图见图 21-23。



(a)

(b)

图 21-23 加垂线与突出线的编辑效果

### 3. 突出线及垂线

(1) Display category range bars，用垂线连接同一分类中不同变量的数值。强调同一分类不同变量值间的差异见图 21-23(a) 垂线长反应两个城市同一月份的温度均值之差。

(2) Projection 突出线栏。选中 Display projection line 项，展示突出线。投影线的作用在于从视觉上区分分类轴上的某值两侧曲线。例如图 21-23(a)是带有垂线的跳跃线图。

① 在 **Start** 下的 **Category** 下拉列表中选择突出线起始点，将光带移到分类轴某个变量值处，突出线将从这个变量值开始。

② **Direction** 下拉列表中选择突出线的方向，即选择分类变量值向前或向后投影。  
图 21-23(b) 是选择从变量值 **Mar** 三月开始，向前加粗的突出线。

## 21.2.7 圆图编辑

### 1. 平面圆图和立体圆图

选中圆图，打开 **Properties** 道具对话框，选择 **Depth & Angle** 深度和角度选项卡，见图 21-24(a)。

(1) **Effect** 图形效果栏：**Flat** 平面效果；**Shadow** 阴影效果；**3-D** 三维效果。

(2) **Angle** 角度栏，对 3-D 效果，仅可移动纵向游标改变纵向观察角度；对阴影效果可以移动纵向和横向游标改变光源方向。根据预览结果确定想要的最佳角度。

(3) **Distance** 距离的操作仅对 3-D 图形有效，确定图形的大小代表距离远近。

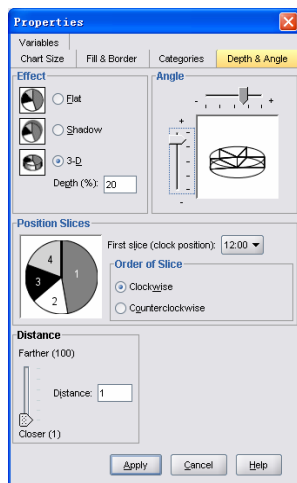
(4) **Position Slices** 栏，确定扇形位置和排列方向。

① **First slice (clock position)**，以钟表盘的方式确定第一个扇面的位置。

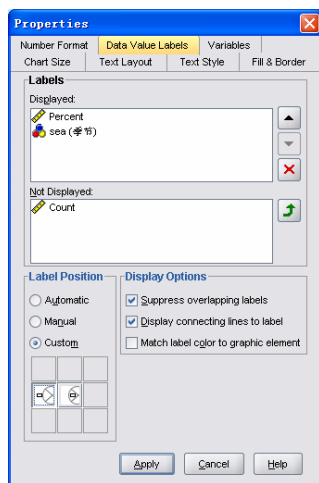
② **Order of Slice** 框，选择圆图中扇形的排列方向。**Clockwise**：顺时针排列；**Counter clockwise**：逆时针排列。

### 2. 数值标签

选择圆图，在右键菜单中选择 **Data Value Labels** 打开 **Properties** 道具窗口数值标签选项卡，见图 21-24(b)。默认的数字标签是百分比。



(a)



(b)

图 21-24 圆图的深度、角度和数值标签选项卡

(1) **Labels** 栏，决定显示什么内容的标签。**Displayed** 框中的是已经显示的标签。**Not**

Displayed 框中是没有显示在图中的, 选中变量拖入 Displayed 框的均可以在圆图中显示。

(2) Label Position 栏, 读者可以自行指定标签的位置。

(3) Display Option 栏, 选择标签显示方式。

① Suppress overlapping labels 选项, 压缩重叠的标签。

② Display connecting lines to label 显示指向图形元素的连接线。

③ Match label color to graphic element 标签与图形元素颜色匹配。

3. 分离扇面的圆图。选择一个扇面, 右键菜单中选择 Explode Slice。

图 21-25 是根据部分独联体国家每季度失业人口数据 data21-03 作出的饼图, 经过编辑后的结果。图 21-25(a)是选择显示百分比和季节的三维圆图。图 21-25(b)是分离第 4 季度扇面的三维圆图。

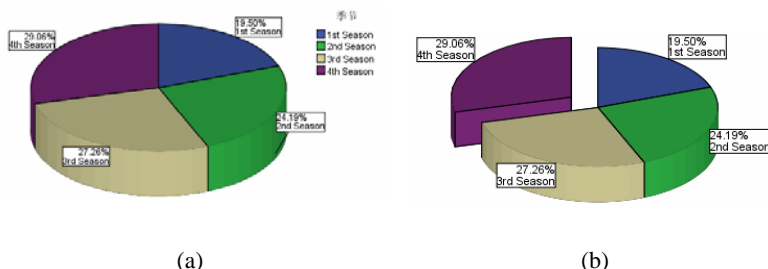


图 21-25 阴影与 3-D 圆图效果

## 21.2.8 散点图的编辑

1. 点样式的编辑。各种类型散点图的道具窗都有 Marker 选项卡。编辑方法都相同。

选中图中的点, 打开 Properties 道具窗口, 选择 Marker 选项卡。可以选择点的类型 (Type)、大小 (Size)、边界线宽 (Border Width) 以及颜色 (Color) 等。

2. 散点图类型在道具窗 Variable 选项卡中进行。在 Element type 下拉列表中选择可以转化成的其他类型的散点图和其他图形。

3. 添加钉线

钉线是指从每个数据点到所选定位置画的线段, 它可用来观察数据点的差异。选中要加钉线的点, 打开道具窗, 选择 Spikes。3-D 和其他散点图的钉线选项卡见图 21-26。

① None, 系统默认无钉线。

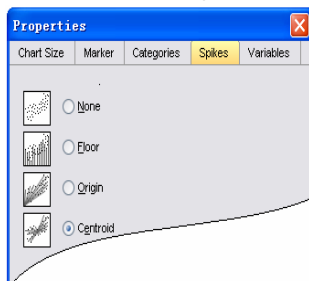
② Floor, 对平面散点图, 钉线为每个数据点到 X 轴的连线。在 3-D 散点图中, 为每个数据点到 XZ 轴平面的连线。

③ Origin, 从每个数据点到原点的连线。

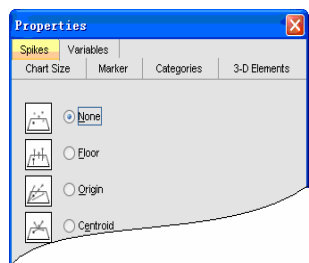
④ Centroid, 从每个数据点到全部数据距心的连线, 距心的坐标是 XYZ 轴上三个变量值的加权平均数, 其中任一变量中有缺失值, 要从计算中剔除。改变轴的刻度将不影响中心点的计算。

图 21-27 是使用 Data21-02 数据做出的简单散点图的钉线 4 种方式的效果。最容易

看出 4 个选择项的含义。图 21-28(a) 是矩阵散点图和矩阵散点图 Floor 钉线效果, (b) 是 3D 散点图的 Origin 钉线的效果。

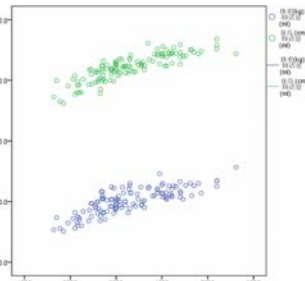


(a) 其他

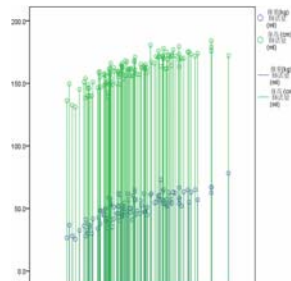


(b) 3-D

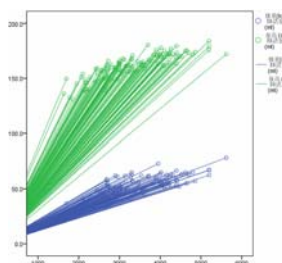
图 21-26 钉线选项卡



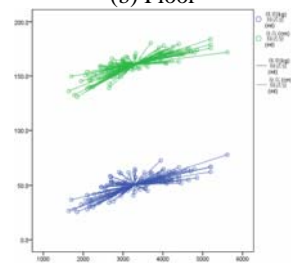
(a) None



(b) Floor

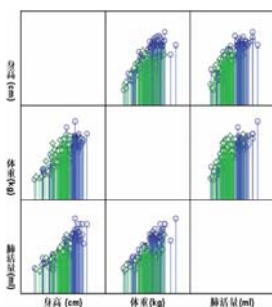


(c) Origin

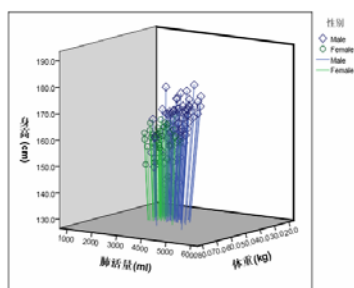


(d) Centroid

图 21-27 简单散点图的四种钉线效果



(a)



(b)

图 21-28 其他钉线举例

#### 4. 拟合线的生成与修饰

(1) 生成拟合线 在编辑器窗口, 选择散点图, 在工具栏或右键菜单中选择 对所有点产生拟合直线; 选择 分类产生拟合直线。图例下方显示线性拟合统计量  $R^2$  Linear, 该值越大拟合得越好。

使用数据 data21-02 做出的肺活量(X轴)和体重(Y轴)散点图和拟合线见图 21-30。

#### (2) 编辑拟合线

在散点图中选中拟合线, 打开 Properties 窗口选择 Fit Line 拟合线选项卡, 如图 21-29

所示。

① 两个复选项：

- **Display Spikes** 显示拟合线到每个点的垂直连线。
- **Suppress intercept** 修改拟合线，使之通过原点。即回归直线方程不包括截距。

② **Fit Method** 栏，选择 5 种另外的拟合方式，可以比较图右侧标出的  $RSq$ （即  $R^2$ ）值，选择最佳拟合效果。如果已知数据趋于线性回归直线、二次回归曲线和三次回归曲线，则可以直接从下面的选项中拟合数据，如果不了解数据集趋于何种曲线，使用 **Loess** 选项对所有数据进行整体拟合。图 21-30 是直线拟合和二次曲线拟合的结果。

① **Mean of Y** 选项：生成一条  $Y$  轴数值的平均线。

② **Linear regression**：线性回归直线，根据最小平方方法，用线性回归直线对散点图中的数据点进行最佳拟合，这是系统默认值方式。

③ **Quadratic regression**：根据最小平方方法，用二次回归曲线拟合散点图中的点。

④ **Cubic regression**：根据最小平方方法，用三次回归曲线对散点图中的点进行拟合。

⑤ **Loess** 局部加权回归散点修匀法。用迭代加权最小平方方法拟合，至少要 13 个点。

- **% of points to fit** 指定用于拟合的数据占总数的百分比，默认值为 50%。
- **Kernel** 选定所需要的核函数。

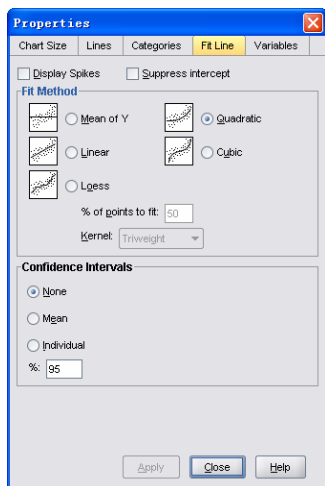


图 21-29 拟合线选项卡

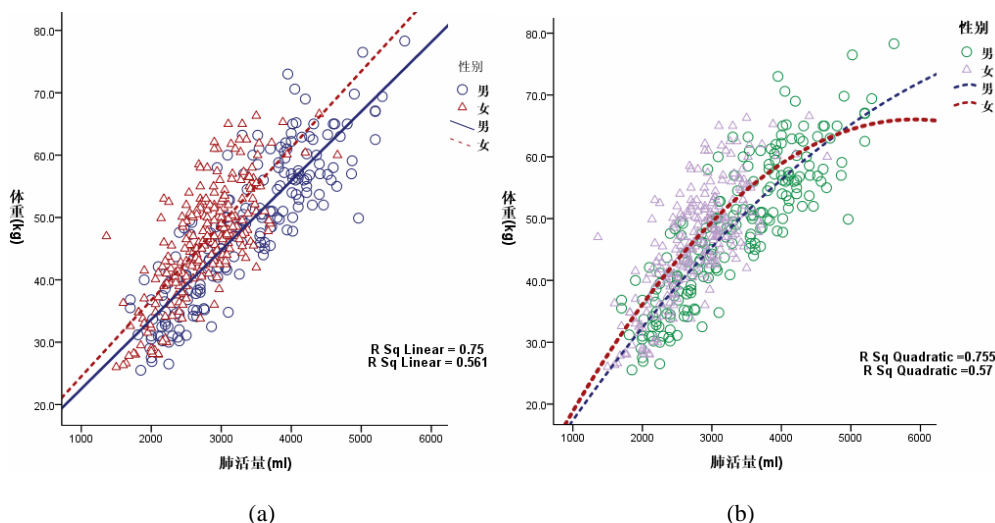


图 21-30 直线拟合与二次曲线拟合与修饰的结果



图 21-30 的  $RSq$  值比较结果：直线拟合的两个  $R^2$  分别是 0.75、0.561，二次曲线拟合的  $R^2$  分别是 0.755 和 0.57。虽然看起来二次拟合比较好，但是差别不大。还应该取更多的观测进行拟合。本例只是取了 451 个观测的结果。

注意，缺失值会对图形有较明显的影响，做图前一定要定义好缺失值，使之不参与绘制图形。

(3) Confidence Intervals 栏，选定拟合线的可信区间。None 不生成可信区间线。Mean 生成平均值的可信区间线。Individual 生成单个观测量的可信区间线。Mean 和 Individual 选项，需要指定可信区间百分比数。

5. 3-D 散点图的编辑

(1) 在右键菜单中选择 3-D Rotation 打开对话框。在对话框中设置不同的水平与纵向数值，如图 21-31(a)、(c)所示。图 21-31(b)、(d)是对应(a)、(c)不同水平与纵向值的 3-D 图实时旋转的效果。

(2) 选择 3-D 散点图，在 Properties 道具窗，选择 3-D Element 选项卡如图 21-32(a)所示。

- ① 选中 Display backplane 显示三维面（背景板）。
- ② 选中 Display wireframe 显示三维框线。
- ③ Distance 栏中调整距图形的距离远近，实际效果是图形的小大，下面显示距离值。

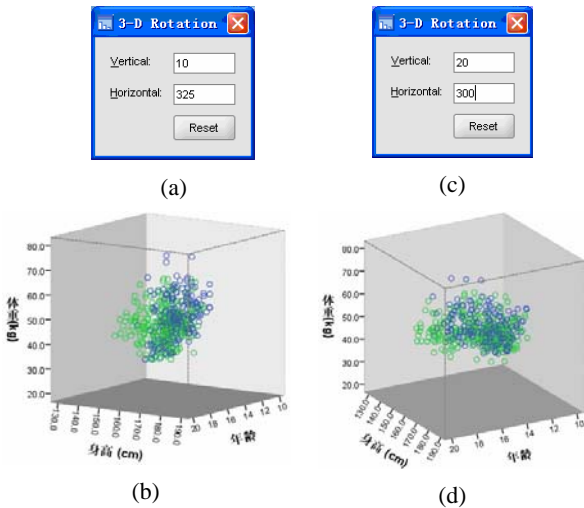


图 21-31 3-D 旋转对话框不同设置的旋转效果

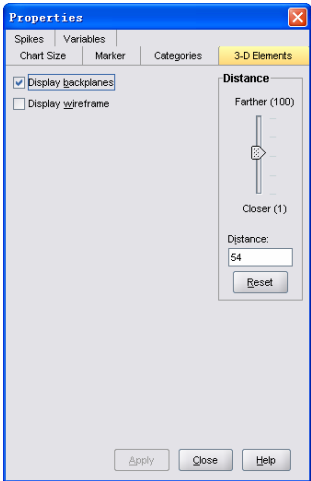


图 21-32 3-D 元素选项卡

21.2.9 文件管理

1. 保存图形模板

读者将生成和完成编辑的图形保存为模板文件，以后在生成新的图形时，调用模板文件，新生成的图形的格式与模板中的图形格式相一致，省去了许多烦琐的编排工作。

按 **File→Save Chart Template** 顺序单击鼠标, 打开保存图形模板对话框, 见图 21-33。可以选择想保存在模版文件中的图形要素。当前图形所有的图形要素都显示在项目框中, 以折叠菜单形式显示在对话框中。可选择保存的种类和细项大致有以下内容:

(1) **All Settings** 保存所有图形元素的设置

① **Layout** 确定了模板图形的版面编排, 包括图形大小、图形的纵横比例、图形各个边框、文本的位置等。

② **Text Content**, 包括图形标题、轴标题、注脚、注释等元素。

③ **Style**, 包括文字格式; 非数据元素样式, 如填充和边界样式、线型、条、点的样式、是否有坐标线、背景板等; 数据元素样式, 针对不同的图形有不同的选择项, 例如 3-D 图的旋转等。

④ **Axis** 栏可选择的有: **Scale Range** 刻度轴的范围, 刻度轴上下空白空间, 刻度轴的数值类型, 主刻度标记和每个主刻度之间次刻度的数量, 显示派生轴; **Scale Settings** 起始位置, 刻度轴的日期或时间的版式, 刻度轴的数值版式; **Category Range** 分类轴左右的空白空间, 合并分类的设置; **Category Settings** 分类值标签的显示方式, 分类标签的排列方式等。

⑤ **Statistics** 栏可选择的有 **Fit and Reference Line** 显示拟合线、参照线、连线和直方图正态曲线; **Scatter Plot Binning** 显示散点图的组化点; **Histogram Binning** 显示直方图的组化图等。

(2) 单击 **Expand All**, 在项目框中显示所有项目。

(3) **Description of Template** 栏, 输入对本模板的描述文字。

单击 **Reset** 按钮, 恢复到选择前的状态, 出现选择要保存的项目。选择完成后单击 **Continue** 按钮打开保存对话框, 选择保存位置和文件名, 单击 **save** 保存完成。

## 2. 应用图形模板

在图形编辑窗口中, 套用某个图形模板, 按 **File→Apply Chart Template** 顺序单击鼠标, 打开 **Apply Chart Template** 应用图形模板对话框, 选择需要的图形模板。

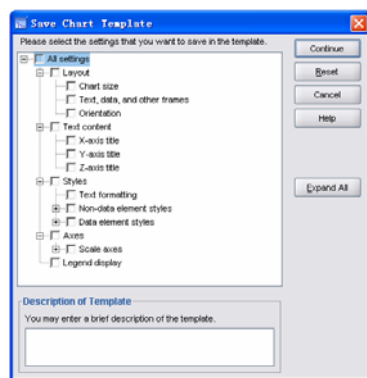


图 21-33 保存图形模板对话框

## 习 题 21

1. 各种图形是由哪些成分组成的? (条形图、圆图、散点图等)

2. data21-03 是世界各地气候、人口状况数据, 做出宗教信仰圆图并修饰之。对世界上不同宗教所占百分比圆图, 做以下调整: 显示每个扇面的文字、数值和百分比; 合并小于 5% 的扇面。

## 附录A 标准化、距离和相似性的计算

SPSS 的许多过程中都使用了距离和相似性、不相似性的计算，例如聚类分析、尺度分析等都会用到。

1. 对于等间隔测度的变量（尺度变量，测度类型为 Scale）计算距离的方法

约定：距离或相似性的公式中  $x$ 、 $y$  均表示  $n$  维空间中的两个点， $x_i$  是  $x$  点的第  $i$  个变量的值， $y_i$  是  $y$  点第  $i$  个变量的值。

(1) Euclidean distance（欧氏距离）。两项之间的差是每个变量值之差的平方和之平方根。

$$EUCLID(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

(2) Squared Euclidean distance（欧氏距离平方）。两项间的距离是每个变量值之差的平方和。

$$SEUCLID(x, y) = \sum_i (x_i - y_i)^2$$

(3) Cosine（cos 相似性测度）。计算值向量间的余弦，值范围是-1~1，用 0 值表明两向量正交（相互垂直）。

$$COSINE(x, y) = \frac{\sum_i (x_i y_i)^2}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$$

(4) Pearson correlation（皮尔逊相关）。计算值向量间的相关，Pearson 相关是线性关系的测度，范围是-1~1。0 值表明没有线性关系。

$$CORRELATION(x, y) = \frac{\sum_i (Z_{x_i} Z_{y_i})^2}{n - 1}$$

(5) Chebychev（切贝谢夫距离）。两项间的距离用最大的变量值之差的绝对值表示。

$$CHEBYCHEV(x, y) = \text{Max}_i |x_i - y_i|$$

(6) Block（布洛克距离）。两项之间的距离是每个变量值之差的绝对值总和。

$$BLOCK(x, y) = \sum_i |x_i - y_i|$$

(7) Minkowski (明可斯基距离)。两项之间的距离是各变量值之差的  $p$  次方幂的绝对值之和的  $p$  次方根。

$$MINKOWSKI(x, y) = \sqrt[p]{\sum_i |x_i - y_i|^p}$$

(8) Customized (自定义距离)。两项之间的距离用各项值之间差值绝对值的  $p$  次幂之和的  $r$  次方根表示。 $p, r$  可以自己指定。

$$MINKOSKI(x, y) = \sqrt[r]{\sum_i |x_i - y_i|^p}$$

## 2. 两个计数变量的不相似性测度的方法

(1) Chi-square measure ( $\chi^2$  测度)。用卡方值测度不相似性。该测度是假设两个集的频数相等进行的卡方检验，测度产生的值是卡方值的平方根。这是系统默认的对计数变量的不相似性测度方法，是根据被计算的两个观测量或两个变量总频数计算其不相似性。期望值来自观测量或变量( $x, y$ )的独立模型，其中  $E(x_i)$  和  $E(y_i)$  为频数期望值。

$$CHISQ(x, y) = \sqrt{\frac{\sum_i (x_i - E(x_i))^2}{E(x_i)} + \frac{\sum_i (y_i - E(y_i))^2}{E(y_i)}}$$

(2) Phi-square measure (两组频数间的  $\Phi^2$  测度)。该测度考虑了减少样本量对测度值的实际预测频率减少的影响。该测度把  $\Phi$  平方除以联合频数的平方根，使不相似性的卡方测度规范化。该值与参与计算不相似性的两个观测量，或两个变量的总频数无关。

$$PH2(x, y) = \sqrt{\frac{\frac{\sum_i (x_i - E(x_i))^2}{E(x_i)} + \frac{\sum_i (y_i - E(y_i))^2}{E(y_i)}}{N}}$$

## 3. 二值变量的距离或不相似性测度的约定

(1) 首先应该明确，对二值变量，系统默认用 1 表示某特性的出现（或发生、存在等），用 0 表示某特性不出现（或不发生、不存在）。

(2) 对二值变量的相似性或不相似性测度都基于一个四格表，见表 A-1。

表 A-1 四格表

	第二特性	
第一特性	发生	不发生
发生	$a$	$b$
不发生	$c$	$d$

如果对观测量进行计算，则对所有“变量对”做一遍四格表。如果对变量进行计算，则对所有“观测量对”做一遍四格表。对每个四格表按所选择的方法进行一次距离参数的计算，这样形成距离矩阵。例如，分析变量  $V$ 、 $W$ 、 $X$ 、 $Y$ 、 $Z$ ，观测量 1 的 5 个变量值顺序为 0、1、1、0、1；观测量 2 的 5 个变量值顺序为 0、1、1、0、0，如下面的表 A-2 所示。

两个事件都发生的有两个变量  $W$ 、 $X$ ，相应的四格表的  $a$  为 2；两个事件都不发生的有变量  $V$ 、 $Y$ ，因此  $d=2$ ；事件 1 发生，事件 2 不发生的是变量  $Z$ ，因此  $b=1$ ；事件 1 不发生，事件 2 发生的没有，因此， $c=0$ 。相应的四格表读者自己可以做出。

表 A-2 例题数据中的两个观测量及对应的四格表

分析变量 观测量号	$V$	$W$	$X$	$Y$	$Z$
1	0	1	1	0	1
2	0	1	1	0	0

		第二特性	
		发生	不发生
第一特性	发生	$a=2$	$b=1$
	不发生	$c=0$	$d=2$

- (3) 对二值数据的相似性或不相似性测度，或二值变量距离测度算法的分类：
- ① 匹配系数的计算，包括：RR、SM、SS1、RT、JACCARD、DICE、SS2、K1、SS3。
  - ② 与条件概率有关的测度包括：K2、SS4、HAMANN。
  - ③ 与预测特性有关的测度包括：Y、Q、LAMBDA、D。
  - ④ 其他距离、相关等测度包括：BEUCLID、BSEUCLID、SIZE、PATTERN、BSHAPE、OCHIAI、SS5、PHI 等。
- (4) 在下面给出的公式中，作为自变量的  $x$ ， $y$  不一定指两个变量。因为用观测量之间的相似性或距离可以进行观测量聚类，那么  $x$ 、 $y$  就指两个观测量；作变量聚类，要计算两个变量的距离或相似性、二值相关，那么这里的  $x$ ， $y$  就是两个变量。
- (5) 另外，表中联合发生的指  $a$ ，联合不发生的指  $d$ ，所有匹配的指  $a+d$ ，所有不匹配的指  $c+b$ ， $n=a+b+c+d$ 。表 A-3 说明匹配系数。
- (6) 按权重和分子分母特征归纳各计算方法如表 A-3。

4. 二值变量的距离或不相似性测度的方法

- (1) Euclidean distance，二值欧氏距离。根据四格表计算  $\text{SQRT}(b+c)$ 。 $b$  和  $c$  表示事件在—项中发生，在另一项不发生的对角单元，其最小值为 0，无上限。
- (2) Squared Euclidean distance，二值欧氏距离平方。计算的是不匹配事件的数目，其最小值为 0，无上限，数值等于  $b+c$ 。
- (3) Size difference，不对称指数，其值范围在 0~1 之间。

$$SIZE(x, y) = \frac{(b-c)^2}{n^2}$$

表 A-3 匹配系数计算中的权重关系及分子、分母特征表

		分子中不包括联合不发生的 $d$	分子中包括联合不发生的 $d$
All matches included in denominator 分母中包括所有匹配的 ( $a$ 、 $d$ )	给匹配与不匹配的权重相等	RR	SM
	给匹配的双倍权重		SS1
	给不匹配的双倍权重		RT
Joint absences excluded from denominator 联合不发生的不包括在分母中 ( $d$ )	给匹配与不匹配的权重相等	JACCARD	
	给匹配的双倍权重	DICE	
	给不匹配的双倍权重	SS2	
All matches excluded from denominator 分母中剔除所有匹配的 ( $a$ 、 $d$ )	给匹配与不匹配的权重相等	K1	SS3

(4) Pattern difference, 不相似性测度。范围为 0~1。根据四格表计算  $bc/n^2$ , 其中  $b$  和  $c$  表示事件在一项中发生, 在另一项中不发生的对角单元。 $N$  是观测量或变量总数。

(5) Variance, 方差不相似性测度。根据四格表计算  $(b+c)/4n$ , Variance 值范围为 0~1。

(6) Dispersion, 是一个相似性指数。其范围为 -1~1。

$$DISPER(x, y) = \frac{ad - bc}{n^2}$$

(7) Shape, 距离测度。范围无上下限。

$$BSHAPE(x, y) = \frac{n(b+c) - (b-c)^2}{n^2}$$

(8) Simple matching, 匹配数与值的总数的比值。它给匹配与不匹配以相同的权重。

$$SM(x, y) = \frac{a+d}{n}$$

(9) Phi 4-point correlation, 皮尔逊相关系数二值模拟, 其值范围为 -1~1。

$$PHI(x, y) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

(10) Lambda 数, 是 Goodman and Kruskal 的  $\lambda$ , 是一种相似性测度。当预测方向同等重要时该系数估计的是用一项预测另一项的误差降低的比例。其值范围为 0~1。

$$LAMBDA(x, y) = \frac{t_1 - t_2}{2n - t_2}$$

其中  $t1 = \text{Max}(a,b) + \text{Max}(c,d) + \text{Max}(a,c) + \text{Max}(b,d)$ ,  $t2 = \text{Max}(a+c, b+d) + \text{Max}(a+d, c+d)$ 。

(11) Anderberg'D 统计量, 类似于  $\lambda$ , 该指数取决于用一项预测另一项 (在两个方向上进行预测) 的误差降低的实际数值。其值范围为 0~1。

$$D(x, y) = \frac{t_1 - t_2}{2n}$$

其中  $t1$ ,  $t2$  定义与 (10) 中的定义相同。

(12) Dice, 该指数中剔除了联合不发生, 给匹配以双倍权重。类似 Czekanowski 或 Sorensen 测度。

$$DICE(x, y) = \frac{2a}{2a + b + c}$$

(13) Hamann, 相似性测度。该指数是匹配数减去不匹配数除以总项数。其值范围是 -1~1。

$$HAMANN(x, y) = \frac{(a+d) - (b+c)}{n}$$

(14) Jaccard, 是一个不考虑联合缺席 ( $d$ ) 的指数。它给匹配与不匹配以相等的权重, 类似相似比。

$$JACCARD(x, y) = \frac{a}{a + b + c}$$

(15) Kulczynski 1, 是联合出现与非匹配数的比。该指数有下界 0, 无上界。理论上对无不匹配的情况 ( $b=0, c=0$ ) 没有定义, 然而, 当值是没有定义的或大于 9999.999 时, 软件赋值给该指数一个武断值 9999.999。

$$K1(x, y) = \frac{a}{b + c}$$

(16) Kulczynski 2, 相似性测度。该指数根据某特性在一项中出现的条件概率给出在其他项中出现的概率。计算该指数时, 每一项作为其他项的预测值时, 各值取其平均数。

$$K2(x, y) = \frac{a/(a+b) + a/(a+c)}{2}$$

(17) Lance and Williams, 根据四格表计算  $(b+c)/(2a+b+c)$ , 其中  $a$  表示与事件在两项中都发生相对应的单元,  $b$  和  $c$  表示事件在一项中发生而在另一项中不发生的对角单元。该测度的值的范围为 0~1。有如我们所知的 Bray-Curtis 非度量系数。

(18) Ochiai, 该指数是余弦相似性测度的二元形式。范围为 0~1。

$$OCHIAI(x, y) = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

(19) Rogers and Tanimoto, 是一个给不匹配的 ( $b$ 、 $c$ ) 双倍权重的指数。

$$RT(x, y) = \frac{a + d}{a + d + 2(b + c)}$$

(20) Russel and Rao, 是内积（点积）的二元形式。对匹配与不匹配都给予相等的权重，是二元相似数据的系统默认方法。

$$RR(x, y) = \frac{a}{a + b + c + d}$$

(21) Sokal and Sneath 1, 给匹配以双倍权重的一种指数。

$$SSI(x, y) = \frac{2(a + d)}{2(a + d) + b + c}$$

(22) Sokal and Sneath 2, 给不匹配以双倍权重的一种指数，且不考虑联合缺席的情况。

$$SS2(x, y) = \frac{a}{a + 2(b + c)}$$

(23) Sokal and Sneath 3, 匹配与不匹配的比。该指数下界为 0, 无上界。理论上, 对无不匹配的情况没有定义。当值为未定义或大于 9999.999 时, 软件给予该指数一个特定常数 9999.999。

$$SS3(x, y) = \frac{a + d}{b + c}$$

(24) Sokal and Sneath 4, 同一匹配状态（某特性出现或不出现）在另一项出现或不出现的条件概率。计算该指数时，每一项作为其他项的预测值时，各项值取其平均数。该指数范围为 0~1。

$$SS4(x, y) = \frac{a / (a + b) + a / (a + c) + d / (b + d) + d / (c + d)}{4}$$

(25) Sokal and Sneath 5, 该指数是正负匹配的条件概率的几何平均数的平方。它独立于项编码。其值范围为 0~1。

$$SS5(x, y) = \frac{ad}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

(26) Yule's Y, 该指数是 2×2 表交叉比的函数，且独立于边际总和，范围为-1~1。有如我们所知的综合系数。

$$Y(x, y) = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

(27) Yule's Q, 该指数是 Goodman 和 Kruskal  $\gamma$  (gamma) 的特殊事件，是 2×2 表交叉比的函数，且独立于边际总和。其值范围为-1~1。



$$Q(x, y) = \frac{ad - bc}{ad + bc}$$

### 5. 对数据进行标准化的方法

① **Z scores**, 把数值标准化到 Z 分数。标准化后变量均值为 0, 标准差为 1。系统将每一个值减去正被标准化的变量或观测量的均值, 再除以其标准差。如果原始数据的标准差为 0, 则将所有值置为 0。

② **Range -1 to 1**, 把数值标准化到 -1 至 1 范围内。选择该项, 对每个值用正在被标准化的变量或观测量的值的范围去除。如果范围为 0, 所有值不变。

③ **Maximum magnitude of 1**, 把数值标准化到最大值为 1。该方法是把正在标准化的变量或观测量的值用最大值去除。如果最大值为 0, 则用最小值的绝对值除再加 1。

④ **Range 0 to 1**, 把数值标准化到 0 至 1 的范围内, 对正在被标准化的变量或观测量的值减去正在被标准化的变量或观测量的最小值, 然后除以范围。如果范围是 0, 将所有变量值或观测值设置为 0.5。

⑤ **Mean of 1**, 把数值标准化到均值的一个范围内。对正在被标准化的变量或观测量的值除以正在被标准化的变量或观测量的值的均值。如果均值是 0, 对变量或观测量的所有值都加 1, 使其均值为 1。

⑥ **Standard deviation of 1**, 把数值标准化到单位标准差。该方法对每个值除以正在被标准化的变量或观测量的标准差。如果标准差为 0, 则这些值保持不变。

## 附录B 数据清单

数据编号	数据名称	出现页码
data02-01.sav	1969—1971 年 美国某银行 474 雇员状况数据	54, 73, 102, 103
data02-02.sav	青少年身高和体重 (1)	58, 60, 61
data02-02a.txt	data02-02.sav 以 ASCII 格式保存 (固定格式, 有列间隔)	60
data02-02b.txt	data02-02.sav 以 ASCII 格式保存 (固定格式, 无列间隔)	60, 61
data02-03.txt	啤酒数据 (固定格式, ASCII 码, 有列间隔)	103
data02-04.txt	啤酒数据 (自由格式, ASCII 码, 有列间隔)	61, 64
data02-05.sav	青少年身高数据 (1)	74, 75
data02-05a.sav	data02-05.sav 数据的简化版	76
data02-06.sav	青少年身高数据 (2)	76
data02-07.sav	青少年身高和体重 (2)	76
data02-08.sav	青少年身高数据 (3)	78, 79, 80
Data02-08a.sav	文件合并结果数据	79
data02-09.sav	青少年体重数据	78, 79, 80
data02-10.sav	237 个人的身高体重数据	84, 101
data02-11.sav	某银行雇员工资和受教育状况 (data02-01.sav 简化版)	85, 86
data02-11a.sav	data02-11.sav 重新编码运行结果	86
data02-12.sav	顾客对 17 种汽车评价	88
data02-12a.sav	data02-12 数据转置结果	88
data02-13-1.sav	学生学号、身高、体重及 3 门课程分数	90
data02-13-1a.sav	data02-13-1.sav 数据结构重建 (ScoreA、ScoreB 和 ScoreC 变量组转换成一个观测量组, 索引变量为顺序值)	92
data02-13-1b.sav	data02-13-1.sav 数据结构重建 (保留固定变量 h、w, 索引变量值为原始变量名)	92
data02-13-1c.sav	data02-13-1.sav 数据结构重建 (不保留固定变量, 索引变量值为原始变量名)	92
data02-13-2.sav	学生身高、体重及文理科分数数据	94
data02-13-2a.sav	data02-13-2.sav 数据结构重建 (以文科和理科成绩各自生成变量) 转换后的数据	94
data02-14.sav	两班学生 3 门课程成绩	95
data02-14a.sav	data02-14.sav 数据结构重建	96
data02-15.sav	期中和期末的 3 门课程成绩	97

数 据 编 号	数 据 名 称	出 现 页 码
data02-16.sav	3 个市场、7 种商品价格	103
data02-17.sav	中国女排档案	68
Data04-01.sav	随机变量分布函数例 1 题结果	131
Data04-02.sav	随机变量分布函数例 2 题结果	131
Data05-01.sav	生日日期型数据	141
Data05-01a.sav	Data05-01 中变量类型转换结果	141
Data05-02.sav	数值型数据及转换成日期型数据的结果	141, 142
Data05-03.sav	字符型数据	142
Data05-03a.sav	Data05-03 字符型转换成日期型数据结果	143
Data05-04.sav	生日数据	143
Data05-04a.sav	字符型日期转换成数值型结果	143
Data05-04b.sav	生日计算年龄的结果	145
Data05-05.sav	从生日提取月份的结果	145
data06-01.sav	不同性别、年龄、婚姻状况的生活方式和首选早餐的调查	134
Data06-02.sav	某公司不同性别经理薪金情况	
data07-01.sav	1991 年美国社会调查	165, 181
Data07-02.sav	1985 年美国 50 个州犯罪记录	168
Data07-03.sav	474 名银行雇员数据	173, 174
data07-04.sav	某公司不同性别经理薪金情况	182
Data07-05.sav	地产评估数据	185
Data07-06.sav	肺癌患者生存时间数据	187
Data07-07.sav	200 例正常人血铅含量	188
Data07-08.sav	150 名 3 岁女童身高数据	189
Data07-09.sav	6400 人生活状况好现代化工具使用调查数据	190
Data07-10.sav	不同质量等级（标准、高级）合金形成温度	190
data08-01.sav	学生身高与体重	196
data08-02.sav	120 名 12 岁男孩身高	202
data08-03.sav	29 名 13 岁男生身高、体重、肺活量数据	207, 208
data08-04.sav	体育疗法对高血压患者疗效	210
data08-05.sav	方便面面饼重量抽检数据	214
data08-06.sav	减肥训练效果数据	214
data08-07.sav	两培训中心标准化考试数据	214

数据编号	数据名称	出现页码
Data08-08.sav	银行雇员工资学历等数据	206
data09-01.sav	不同饲料比较数据	221, 226
data09-02.sav	不同细菌对三叶草含氮量的影响	229
data09-03.sav	四个种系雌鼠子宫重量	240
data09-04.sav	药物对红细胞增加作用	243
data09-05.sav	不同土壤对甜菜产量影响	246
data09-06.sav	镉作业工人肺活量与年龄、接触时间数据	249
data09-07.sav	教育心理学研究数据	251, 287
data09-08.sav	1481 个心梗患者的数据	257
data09-09.sav	刺激与反应时测量数据	272
data09-10.sav	四种药物对某生化指标的作用（重复测量设计）	276
data09-10a.sav	data09-10.sav 数据结构重建	279
data09-11.sav	两种记忆方法的比较	279
data09-12.sav	航空公司、零售业、旅馆业和汽车制造业评定数据	291
data09-13.sav	银行雇员基本情况和工资数据	291
Data09-14.sav	三种麻醉诱导方法在不同时相测量的收缩压变化	291
data10-01.sav	安徽省国民收入与城乡居民储蓄存款余额	295
data10-02.sav	474 个银行雇员数据	296, 317
data10-03.sav	前 10 名运动员长拳和长兵器两项得分	298
data10-04.sav	四川绵阳地区中山柏生长与自然环境关系资料	307, 314
Data06-05.sav	身高、体重、肺活量数据	317
Data10-06.sav	太阳镜销售情况	317
data11-01.sav	汽车数据	335
data11-02.sav	乳腺癌细胞淋巴转移数据	343
data11-03.sav	1992 年美国总统大选得票结果	352
data11-04.sav	鱼藤酮杀虫剂浓度与杀虫量数据	359
data11-05.sav	美国 1790—1960 年人口数据	366, 367
data11-06.sav	教学实验数据	369, 371
data11-10.sav	某企业 1987—1998 年科研经费与经济效益数据	373
data11-11.sav	某商场 1989—1998 年商品流费用率与商品零售额	373
Data11-12.sav	电流刺激农场动物的实验数据	373

数 据 编 号	数 据 名 称	出 现 页 码
Data11-13.sav	474 名银行雇员工资数据	329
data12-01.sav	300 次掷一颗六面体实验观测结果 (原始录入方式)	376, 377
data12-01a.sav	data12-01.sav (频数录入方式)	377
data12-02.sav	100 名健康成年女子血清总蛋白含量 (原始录入方式)	377, 396
data12-02a.sav	data12-02.sav (频数录入方式)	377
data12-03.sav	31 次掷一枚比赛挑边器实验观测结果 (原始录入方式)	379
data12-03a.sav	data12-03.sav (频数录入方式)	379
Data12-04.sav	掷硬币数据	381
data12-05.sav	质点数与时间间隔数据 (原始录入方式)	383
data12-05a.sav	data12-05.sav (频数录入方式)	383
data12-06.sav	两种安眠药效果对比	385
data12-07.sav	4 种不同操作方法优等品率实验数据	387
data12-08.sav	锻炼前后晨脉比较	389
data12-09.sav	顾客对 3 种款式的衬衣的喜爱程度	390
data12-10.sav	设备进行寿命试验, 记录 10 次无故障工作时间	396
data12-11.sav	监听装置接收信号实验数据	396
data12-12.sav	两个地点的地表土壤 pH 值	396
data12-13.sav	某种药物治疗前后血压变化	396
data12-14.sav	20 个村民对 4 位候选人满意度调查	396
data13-01.sav	汽车性能指标与销售数据	404
data13-02.sav	10 名游泳运动员的三项测试	416, 417
data13-02a.sav	data13-02.sav (作为初始聚类中心的种子数据)	417, 418, 421
Data13-02b.sav	聚类结果的类中心--新种子数据文件	417, 418
data13-03.sav	20 种啤酒数据	428, 437
data13-04.sav	有关学生 10 个测验项目数据	439
data13-05.sav	鸢尾花分类与特征	452, 461
data13-06.xls	74 个国家人口出生率和死亡率 (Excel 格式数据)	471
data13-07.xls	标枪运动员等级成绩数据 (Excel 格式数据)	471
Data13-07a.sav	标枪运动员等级. 素质成绩数据	471
Data13-07b.sav	待判等级的标枪运动员素质数据	471
data14-01.sav	标准大城市人口调查区的 5 个经济学变量数据	478, 484

数据编号	数据名称	出现页码
data14-01a.sav	data14-01.sav (带有因子分数变量)	487, 488
data14-02.sav	顾客对车型偏好的研究调查	492
data14-02a.sav	data14-02.sav (带有因子分数好聚类结果)	493, 494
Data14-02b.sav	data14-02.sav (带有因子载荷)	495
data14-03.sav	10 个省份主要消费支出比例	504
data14-04.sav	某医院 3 年的治疗与经营数据	509
data14-05.sav	31 个省市自治区各种经济类型资产占总资产比重	509
data15-01.sav	研究运动员意志品质的调查	515, 516
data15-02.sav	顾客对饮料相似性感受的调查	520
data15-03.sav	受试者对牙膏品牌的认识数据	522
Data16-01.sav	酸奶调查设计文件	534, 536
Data16-02.sav	带有两个模拟观测的酸奶调查设计数据	535
Data16-03.sav	地毯清洁器的调查设计	550, 551, 552
Data16-04.sav	地毯清洁器调查数据	551, 552
data17-01.sav	某公司 1973—1999 年的销售额	561
data17-02.sav	85 地区宽带供应商 1999 年 1 月—2003 年 12 月服务用户数量	563, 576, 577, 581,
data17-03.sav	Data17-02 补进 2004 年 1—3 月份数据	581
data17-04.sav	某公司 1986—1997 年各季度商品销售数据	584, 586
data17-05.sav	某国际航线 1949 年 1 月—1960 年 12 月的旅客数据	589
data17-06.sav	1989 年 1 月—1998 年 12 月三种男女服装产品销售情况	591
data17-07.sav	某邮购公司 1989 年 1 月—1998 年 12 月服装销售及宣传等数据	592
data18-01.sav	公务员的营养、运动及心理调查数据 1	596, 602, 609, 610
data18-02.sav	公务员的营养、运动及心理调查数据 2	598, 610, 612
data19-01.sav	不同饮食下 90 只老鼠的无肿瘤时间	618
data19-02.sav	58 例肾上腺样瘤患者不同治疗方式下的生存时间	623
data19-03.sav	137 位肺癌患者生存时间	629
data19-04.sav	3 期和 4 期黑瘤患者的数据	631
Data19-05.sav	63 例患者的生存时间、结局及影响因素	631
data20-01.sav	我国 12 个城市 1985—1994 年每城市月平均气温	632, 641, 642, 646
data20-02.sav	数据同 Data20-01.sav, 不同文件结构	632
data20-03.sav	1985—1994 年上海月平均气温	632

数 据 编 号	数 据 名 称	出 现 页 码
data20-04.sav	1988—1992 年世界各种饮料产量	634, 637
data20-05.sav	451 青少年生理形态数据	637, 650, 651, 654, 655, 656, 664
Data20-06.sav	不同时期不同性别毕业生初始薪酬	638
data20-07.sav	1950—1985 年我国国防、经济支出	640
data20-08.sav	1996.4.1—1996.4.19 上证所地产类股票价格	642
data20-09.sav	1996.4.1—1996.4.19 上证所北京地区股票价格	643, 644, 646
data20-10.sav	1996.4.1—1996.4.19 上证所几支工业和商业股票价格	644
Data20-11.sav	某年部分独联体国家失业人口数据	647, 648, 649
Data20-12.sav	银行雇员工资数据	650, 651, 652, 663
data20-13.sav	150 名 3 岁女童身高	657
Data20-14.sav	200 名正常人血铅含量	657
data20-15.sav	某刀具厂切削刀次品件数分类计数	668
data20-16.sav	各国医疗保健从业人数	668, 670
data20-17.sav	汽车空调蒸发器故障分类及计数	669
data20-18.sav	各国制造加工业雇佣人数 100 人以上工厂数量	670
data20-19.sav	男女各年龄司机每百万公里伤亡和非伤亡事故数据	671
data20-20.sav	1988 年 6 月 1 日—30 日每日早中晚三班电解工序的电解效率(1)	672, 674
data20-21.sav	1988 年 6 月 1 日—30 日每日早中晚三班电解工序的电解效率(2)	672
data20-22.sav	某搅拌站实测混凝土坍落度数据	673
data20-23.sav	某种小螺钉检测数据	673, 676, 675, 676
data20-24.sav	某医院每月出现危急外科手术例数	674, 676
data20-25.sav	某轧钢厂生产的 6mm±0.4mm 厚度钢板测试记录	675
data20-26.sav	某构件厂产品质量数据	675, 676
data20-27.sav	抽样数不等的小螺丝检测数据	675, 677
data20-28.sav	世界人口数据	677
Data21-01.sav	我国 12 城市平均气温	680
Data21-02.sav	451 名青少年体质数据	693, 694
Data21-03.sav	某年部分独联体国家失业人口数据	693, 697

## 参 考 文 献

1. George A. Morgan, Orlando V. Griego. Mahwah. Easy use and interpretation of SPSS for Windows: answering research questions with statistics. NJ Lawrence Erlbaum, c1997
2. Duncan Cramer. Introducing statistics for social research: step-by-step calculations and computer techniques using. SPSS. London Routledge, 1994
3. SAS/BASE guide for Personal Computer. SAS Institute Inc, 1988
4. SAS/STAT Guide for Personal Computer. SAS Institute Inc, 1988
5. SPSS Base 7.5 for Windows user's guide. SPSS Inc, 1997
6. SPSS graphics. SPSS Inc, 1985
7. Marija J. Norusis. SPSS professional statistics 6.1. Chicago IL, 1994
8. Naresh K Malhotra. Marketing Research. 市场调研. 北京: 清华大学出版社, 1998年第1版
9. 卢纹岱, 金水高. SAS/PC统计分析实用技术. 国防工业出版社, 1996
10. 高惠璇, 张庆峰等编译. SAS系统与市场调查数据分析. 北京大学概率统计系, 1997
11. 吴明隆. SPSS系统应用务实. 北京: 中国铁道出版社, 2000年第1版
12. 汪贤进. 常用统计方法手册. 杭州: 浙江人民出版社
13. Douglas M Bates. 非线性回归分析及其应用. 北京: 中国统计出版社, 1998
14. D.A.Ratkowsky. 非线性回归模型. 南京: 南京大学出版社, 1986
15. 郝德元. 教育与心理统计. 北京: 教育科学出版社, 1982
16. Elisa T Lee. 生存数据分析的统计方法. 北京: 中国统计出版社
17. 孙尚拱. 实用多变量统计方法. 北京: 中国医科大学与中国协和医科大学联合出版社, 1990
18. 吴国富. 实用数据分析方法. 北京: 中国统计出版社
19. 袁淑君. 数据统计分析——SPSS/PC<sup>+</sup>原理及其应用. 北京: 北京师范大学出版社, 1995
20. 周兆麟. 数理统计学. 北京: 中国统计出版社, 1987
21. 张元. 田间实验与生物统计. 沈阳: 东北师大出版社, 1986
22. 贾宏宇. 统计辞典. 上海: 上海人民出版社, 1986
23. 郑家亨. 统计大辞典. 北京: 中国统计出版社
24. David F Freedmen. 统计学. 北京: 中国统计出版社, 1997
25. 胡学锋. 统计学. 广州: 中山大学出版社, 1999
26. 黄德霖. 统计学. 北京: 人民日报出版社, 1988
27. 杨树勤. 卫生统计学. 北京: 北京人民卫生出版社, 1993
28. 胡良平. 现代统计学与SAS应用. 北京: 军事医学科学出版社, 1996



29. 方积乾. 医学统计学与电脑实验. 上海: 上海科学技术出版社, 1997
30. 史秉璋. 医用多元分析. 北京: 北京人民卫生出版社, 1988
31. 金丕焕. 医用统计方法. 上海: 上海医科大学出版社, 1992
32. 贾怀勤. 应用统计学. 北京: 对外贸易教育出版社
33. S Weisberg. 应用线性回归. 北京: 中国统计出版社, 1998
34. 吴辉. 英汉统计词汇. 北京: 中国统计出版社, 1987
35. 杨树勤. 中国医学百科全书——医学统计学. 上海: 上海科学技术出版社, 1985
36. 最新质量统计技术及其应用. 北京: 机械工业出版社
37. 柯惠新, 丁立宏编著. 市场调查与分析, 北京: 中国统计出版社, 2000年3月
38. 郑日昌, 蔡永红, 周益群. 心理测量学. 北京: 人民教育出版社, 1999年9月第1版
39. 袁淑君, 孟庆茂. 数据统计分析——SPSS/PC\*原理及其应用. 北京: 北京师范大学出版社, 1995年2月第1版
40. 谢小庆. 信度估计的系数[J]. 心理学报, 1998(30). 2: 193—196
41. 侯杰泰. 信度与度向性: 高Alpha量表不一定是单向度[J]. 教育学报(香港), 1995(23), 1:142
42. R.A.Johnson, D. W.Wichern著, 陆璇译实用多元统计分析(第四版), 北京: 清华大学出版社. 2001年4月
43. 孙振球, 徐勇勇. 医学统计学. 北京: 人民卫生出版社, 2002年8月
44. Naresh K.Malhotra. 市场调查(第二版). 北京: 清华大学出版社, 1998年8月第1版
45. David Freedman. 统计学. 北京: 中国统计出版社, 1997
46. S.Weisberg. 应用线性回归. 北京: 中国统计出版社, 1998
47. Douglas M.Bates. 非线性回归及其应用. 北京: 中国统计出版社, 1997
48. 胡学锋. 统计学. 广东: 中山大学出版社, 1999
49. SPSS Advanced Models 10.0. USA: SPSS Inc, 2000
50. SPSS Regression Models 10.0. USA: SPSS Inc, 2000
51. 阮桂海等. SPSS for Windows高级应用教程. 北京: 电子工业出版社, 1998
52. 孙明玺. 预测和评价. 浙江: 浙江教育出版社, 1986
53. 于秀林, 任雪松. 多元统计分析. 北京: 中国统计出版社, 1998
54. 徐国祥. 统计预测和决策. 上海: 上海财经大学出版社, 1998
55. George E.P.Box. 时间序列分析预测与控制. 北京: 中国统计出版社, 1997
56. 吴喜之. 非参数统计. 北京: 中国统计出版社, 1999
57. 张建华, 王健等译. 商务与经济统计(第七版). 机械工业出版社, 2000年4月第1版
58. 陈鹤琴, 罗明安译. 例解商务统计. 北京: 清华大学出版社